COMMENT AND DEBATE
Social class differences in early cognitive development

(Received April 2015 Revised June 2015) http:// dx.doi.org/10.14301/llcs.v6i3.361

John Bynner

Few pieces of recent longitudinal research have had as much influence in United Kingdom policy circles as Leon Feinstein's analysis of 1970 birth cohort study data (reported in 2003) on cognitive development assessed at ages 22 months, 42 months five years and ten years. Breakdown by social class of the test performance data demonstrated that infants of superior cognitive ability in the first assessment from working class backgrounds showed relative decline with age in test performance compared with their middle class counterparts, who while starting from an inferior position, subsequently overtook the working class group. The findings were embodied in what became a famous graph showing this crossover and consequent reversal of predicted life chances. They pointed to substantial obstacles to social mobility and were an important factor in the policy response of major pre-school educational interventions such as the Sure Start programme introduced by the Labour Government to reverse the trend, which attracted support from across the political spectrum.

Subsequent re-analysis by John Jerrim and Anna Vignoles challenged the existence of the crossover as a statistical artefact attributing it to the well-known phenomenon of 'regression to the mean'. The consequence was a cooling off of support for the intervention policy directed at strengthening working class children's early cognitive performance. This shift included termination of the Sure Start programme by the new Coalition Government (Conservative and Liberal Democrat) that took office in 2010. Subsequent research has qualified the picture further raising issues on a number of methodological and substantive fronts – especially the need to give more attention to measurement error in such work and for a more nuanced interpretation of such longitudinal research results. Social class disadvantage in cognitive development is well established but the relative loss of competence developmentally needs to be treated with caution.

In an opening paper Leon Feinstein reviews the methodological criticism of his original research. The points he raises are then debated in commentaries by John Jerrim and Anna Vignoles, Harvey Goldstein and Robert French, Elizabeth Washbrook and RaeHyuck Lee and Ruth Lupton. Leon Feinstein's response to these will be published in the next issue of the journal.

Opening paper by

Leon Feinstein Early Intervention Foundation, UK
leon.feinstein@eif.org.uk

Social class differences in early cognitive development and regression to the mean

Introduction

In April 2011 the then new Coalition Government published its social mobility strategy (HM Government, 2011). As a minor reference within the overall document, figure 2 of Feinstein (2003) was reproduced on page eight as a reference to the claim that "Bright children from poorer families tend to fall back relative to more advantaged peers who have not performed as well."

This claim in the strategy brought an immediate response in a press release from Professor Daniel Read (2003) of Warwick Business School[1] claiming

that:

> I am very worried that this graph is being used to shape policy when in fact many statisticians will instantly see that it simply replicates a statistical trap or artefact called 'regression toward the mean'. The apparently shocking pattern of results in the graph is simply what statisticians would expect when you measure extremes of performance in two populations of differing ability. The Feinstein graph is constructed... with undue emphasis on extreme results.

Simultaneously, Jerrim and Vignoles (2011), published a paper on the "use (and misuse) of statistics," undertaking a series of simulations and new analyses based on a model assuming "true abilities" with pre-determined social class mean gaps, which shows that under reasonable assumptions the pattern in the chart could result from regression to the mean.

The 1970 Cohort Study is no longer such a significant source of information on the degree of British inequality in contemporary childhood. We have now much larger and more recent studies such as the Millennium Cohort Study, the first major, public United Kingdom (UK) birth cohort study since the 1970 Cohort and the Life Study at University College London (UCL) with an intended sample of 100,000 babies and a wide range of developmental, health epigenetic and neuro-scientific observations.

But the 1970 Cohort data in general are still of considerable interest and many of those who may have quoted the graph in the past may have been disappointed to learn that they had been so badly misled so I am grateful to the editors of this journal for the opportunity to respond to the critique and to raise a handful of questions for further debate. As the quotations above make clear, there is a wider debate both in learned journals and on the pages in the national newspapers. They do not operate by the same rules. I am grateful to the editors for inviting this paper in a special comment and debate section of the journal that is also about the relationship between research, policy and practice.

Some in the policy world have used the graph without understanding it, though it is perhaps hard to see this an issue specific solely to this chart. Some of those who have used it have understood its weaknesses but found it informative, some have had no idea and are horrified it is all so complicated. Few will have much mind to it. Without wishing to reinstate the graph as a "killer chart"[2] I would like, in this paper, to correct some of the mis-representations that have been suggested and suggest some issues for further discussion.

The graph was based on a relatively small sample from a time and place that is, with due respect to its members, now distant. The Britain of today is transformed in terms of ethnicity and the way inequality is experienced. Much larger datasets are available with much larger samples and more consistent measurement across a broader range of aspects of cognitive development. New methodologies are available. So my argument is not that anyone should return to this chart as the way to model and measure the interaction of a distal measure like class and tests of cognitive ability through childhood, but that there are a range of ways of modelling these data, recognising the importance of the measures used, the age at which children are tested, how different models lead to different tests of this social level interaction and how this plays out in different times and places.

I set out below the basic facts of the graph and then discuss in more detail what I think the graph means and raise some questions about meaning and inference, in particular with regard to the definition of true ability and the difference between average, macro-social phenomena and the lives of individuals. The first section sets out first the charts and then their source. The second section describes some of the challenges for inference that it has raised. The third section concludes with reference to these themes.

The chart

Figure 1: Average rank of test scores at 22, 42, 60 & 120 months, by socioeconomic status (SES) of parents

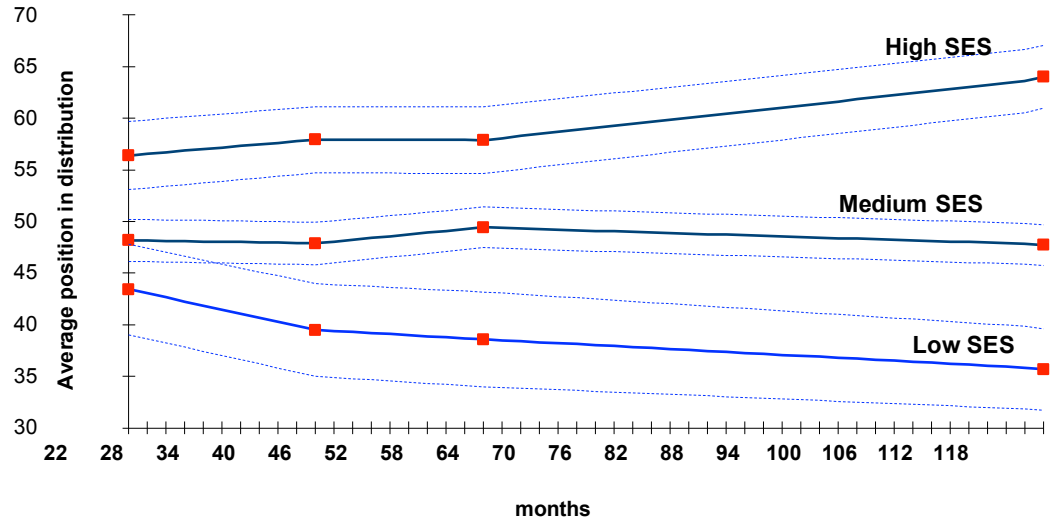


Figure 2: Average rank of test scores at 22, 42, 60 & 120 months, by SES of parents and early rank position

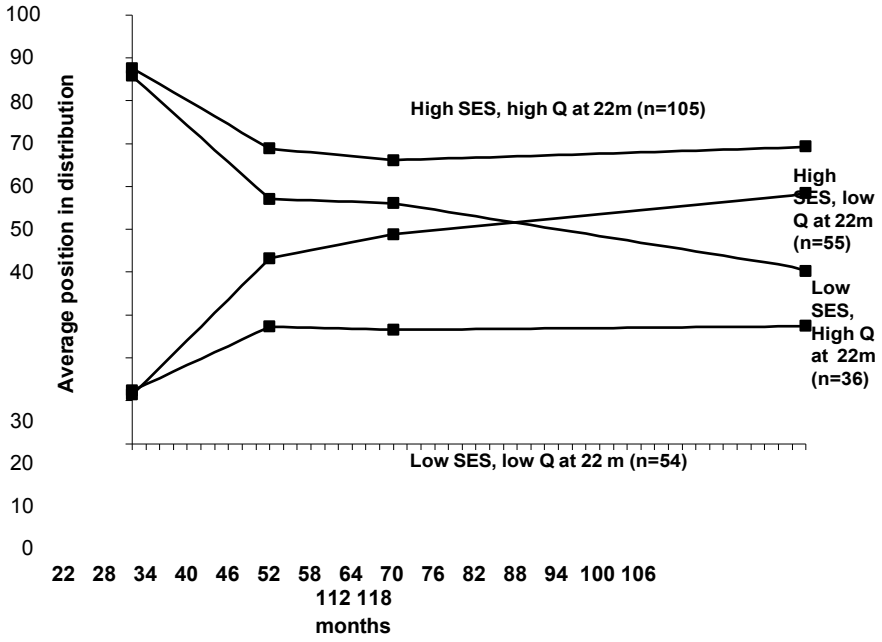


Figure 1 in Feinstein (2003) reports the mean relative positions of children in the 1970 British Cohort Study from different social class groups in age appropriate and hence very different tests of cognitive development at four ages in early and middle childhood (ages 22 months, 42 months, five years and ten years). The explicit rationale of the paper was to test the extent to which average gaps in cognitive development between children from different types of family background were evident before children started school. The difficulty of testing this and of comparing gaps at different ages seemed to centre on the difficulty of finding comparable tests through the complex qualitative developmental changes that occur in early child development. The innovation of the paper was in finding a coherent way to simplify and address the problem. It made possible a cursory, uni-dimensional study of development at the level of average groups of children.

Figure 2 indicates some of the interaction between family background and early ability scores, at particular points of the distribution of early scores.

The main rationale of the paper and of initial discussion of it was of the finding of early socioeconomic status (SES) gaps in figure 1. Figure 2 subsequently became used by me and others in public debate, showing that the children from high SES backgrounds who scored poorly on the age 22 month tests had a higher mean score at 10 years than the low SES children who scored well in the early tests. The shift in relative mean position occurs between the age of five and ten. This always seemed to me remarkable, though recognising that it may be due in untested proportions to measurement error, context, gene*environment interactions, gene*environment correlations and cultural bias. As I said at the time, the graph does not and cannot resolve the question of explanation but it does describe a very common pattern in UK data.

Methods and measurement

Sample

The 1970 Cohort Study was a representative sample based on children born in the first week of April 1970 undertaken initially by the Department of Child Health at Bristol University and taken on to adulthood by the International Centre for Child Studies, City University and then the Institute of Education. Data was collected from members of the cohort studies and from members of their households, through a range of tests and interviews at different ages, as well as through interviews with teachers and medical officers.

The full sample comprising 17,196 children were studied at birth, of whom 13,135 were picked up at age five years and 13,871 at ten years. For the 2003 paper, data were also drawn from the 22 month and 42 month sub-samples of the study. These studies initially comprised 2,457 individuals, of which half were selected from the full sample because at risk of foetal malnutrition and half as a random control group. Attrition and non- response provides a sample of 1,292 children providing test score data and a measure of SES at ages 22 months, 42 months, five years and ten years. This degree of sample loss represents a tremendous achievement by the study team. It is not unproblematic if attrition is non-random such that the remaining sample is no longer typical of the population it is intended to represent. The kind of non-randomness required to cause bias to the general indications of figure 1 and 2 would be one in which children most likely to have been omitted were ones for whom on average the relationship between SES and the development of test scores was different than for the included children.

There is no evidence of this. The paper did have an eye to the differences between children in the control group and those who were selected for the sub-sample because of a distal concern about foetal development. Both issues of sample selection and attrition merit further work. If the paper were to be published today referees would require a greater focus on the handling of missing data than was the case before Little and Rubin (2002) and others had created software and approaches to handling missingness and a greater appreciation of its importance.

Socioeconomic status (SES)

The measures of occupation deposited in the ESRC Data Archive that holds the 1970 Cohort Study were developed by the study team to classify the children by social class using the Registrar-General's Social Classes, which were introduced in 1913 and were renamed in 1990 as Social Class based on Occupation. This was then further aggregated in the 2003 paper into three groups comprising 307 children in a high SES group, 814 in a middle SES group and 171 in a low SES group. There were two further simplifications. Firstly, only one measure of

SES was used for each child, rather than allowing it to change through childhood as family circumstances evolved. The birth measure was used where this was available. Second, the measure was an aggregation of the SES of the male and female adults in the household, in most cases the biological parents. The majority of the mothers of the children in the 1970 Cohort were not working and for these children SES was categorised on the basis of the father's occupation. Where both parents were working and had occupations in different SES groups, high or low SES dominated in the categorisation, such that a child in a household with a high SES father and middle SES mother would be categorised as high SES and vice versa. The very small number of high/lows were categorised as middle.

The intention was to create a simple indicator of occupational skills and access to earnings and status in the economy and society of the day. The broad pattern of results was found to be very similar if groups were constructed on the basis of parents' education levels (Feinstein 2003). The intention was not to indicate that these groupings reflected common genetic inheritances or to indicate anything causal about social class nor to test hypotheses about the nature of social class. Other ways of classifying the children based on different ways of reflecting aspects of family backgrounds such as the distinct contributions of work situation and market situation (see e.g. Erikson 1984) or different ways of reflecting mothers' and fathers' contributions to each would have been possible and equally of interest. These are broad averages based on the available data on the occupation of parents.

Cognitive performance

Comparing measures of cognitive performance at different ages in childhood is made particularly difficult by the qualitative shifts in development that transform the meaning of cognitive capability as children mature. At each age in these data a different set of tests are taken by the children because the tests span the period from early to middle childhood through which considerable shifts in the meaning, nature and measurable manifestation of cognitive development takes place. Piaget (1952), for example distinguishes the sensorimotor stage, from birth to age two in which infants seek and find knowledge through sensory experiences and manipulation of objects; the preoperational stage, from age two to about age seven in which pretending and play are evidently essential to learning; and the concrete operational stage, from age seven to 11 in which logic becomes more routine if at times rigidly applied. In Piaget there is also a formal operational stage, which begins in adolescence and spans into adulthood with an increase in logic, the ability to use deductive reasoning, and an understanding of abstract ideas. There are of course many other models of the nature of developmental change in this period but it is not disputed that very fundamental qualitative change in behaviour and capability occurs through early childhood.

The 22 months tests comprised cube stacking, measures of personal development, measures of language use and a drawing task. The information was collected by a health visitor recruited by the Department of Child Health at the University of Bristol during a visit to the home or other residence of the sample children[3]. The personal development measures were a set of requests from the health visitor to the child such as to point to her nose or her eyes. The cube-stacking test was designed as a test of motor ability, a precursor of later capabilities such as intelligence as well as of physical dexterity. The measures of language use concerned the child's ability reported by the mother to say "ma-ma" and "da-da" and associate the words with the appropriate persons.

The measures as a whole were less well standardised than measures available in the newer UK cohort studies (Chamberlain and Davey, 1976), but merit further study. At age ten years the tests were of maths and reading and the British ability scale test of IQ.

Therefore these measures reflect the transition from early public language use as a 22-month-old child responds to the request from a health visitor to point to her nose to the experience of the age ten child sitting a maths test in a classroom. The relationship between the different features of cognitive and non-cognitive development that lead from the 22 month child pointing to her nose when asked to and the girl sitting the exam are only partially understood but it is clear that there is no simple linear relationship between specific domains of development in early childhood and equivalent domains in adolescence or adulthood.

At 42 months there were tests of what were called in Feinstein (2003) "counting" (use of cubes, such as counting the number of cubes placed on a

table) and "speaking" (correctly naming pictures such as of a car) and a copying designs test, all conducted by trained researchers or health visitors in the home. At age five years there was a test of vocabulary, a copying designs test and a human figure-drawing test (Feinstein, 2003). The age 42 month speaking and counting tests are equal in the prediction of the age ten maths and reading scores with little domain continuity spanning across early to middle childhood (Feinstein, 2003). Not surprising as the tasks all involve elements of language, communication, motor skills and attention, amongst other capabilities.

More and better modelling of this issue is possible now in multiple datasets, not least the Avon Longitudinal Study of Parents and Children which has annual measurement across a much better set of measures of cognitive and other development than were possible in the 1970 Cohort Study.

It is important to emphasise that in order to derive the best possible signal from the available data, the measure used in the graph at each age is not any single test but rather the ordinal position of the children at each age in a weighted average of the scores available at that age. The particular form of weighting is the first principal component, chosen to maximise the variance in the weighted index. Particular tests count higher in the weighting if they add more unique information than other tests. The results are robust to the use of other weighting schema such as regression weights taken from regressing the age ten tests on the earlier tests. Therefore, although the measures used change through development the dependent variable itself is always the relative achievement of the sample children at each age in the age appropriate measures of cognitive development available in the 1970 Cohort Study.

The implications of this are that the dependent variable is ordinal position and has meaning only in this relative sense. It does not measure achievement on any specific test but is an estimate of relative cognitive capability in the available age-appropriate measures. It is therefore different to the repeat measures test that Jerrim and Vignoles (2013) use in their analysis of the Millennium Cohort Study.

Descriptive findings

Figure 1 reports the average (mean) relative positions of children classified by the three-fold categorisation of social class on the single measure of relative ability drawn from the range of age appropriate but different tests at each of the four ages.

Figure 2 reports the mean scores of children from different social class groups at the four ages. Critically, it classified children according to their scores on the first set of tests at age 22 months and considers also the mean scores for children depending on their rank in the early scores.

The early scores are particularly unstable. As I show in the original paper they contain sufficient real information about early development to predict final educational qualifications achieved but the correlation is only just statistically significant and weak. Therefore, as I also show in the original paper, it is not surprising that children's scores subsequently move around a great deal.

Subsequent discussion has focussed on the meaningfulness of the crossover in figure 2, regression to the mean, the appropriateness of classifying children to ability groups at 22 months and the focus on arbitrary quartile groups.

Inference and regression to the mean

When the graph was first presented at an econometrics seminar at UCL issues were raised about measurement error and causality. Then and since there have been debates about regression to the mean, a notion whose original use was by Galton (1886) who discussed the issue in terms of the tendency of the individual deviation from the mean in height to be larger in one generation than the next, so that tall parents will tend to have slightly less tall children. The degree of randomness in a variable through measurement error or chance will mean that the deviation from the mean in one period is that much less likely to be replicated in the next.

In this instance regression to the mean refers in part to the statistical property by which because of misclassification bias in the early groupings, those who appeared to do well early on will have a tendency to lower scores in subsequent tests. Conversely, those who score badly early on will have a tendency to better scores. This was shown by Tu and Law (2010) to be a fatal problem for interpretation of the chart as the outcomes for those from different social class groups with different "true" ability.

Based on their resulting modelling Jerrim and

Vignoles (2011; 2013) have shown how regression to the mean plays out in the kind of data shown in the chart and also used the phrase in other ways. Using simulations, they show that the pattern observed in the chart can be substantially reproduced as the result of regression to the mean of various kinds, including both error in measurement and hence classification to high and low groups and differences in what is tested at different ages.

They reference Nick Clegg's use of the phrasing that: "By the age of five, bright children from poorer backgrounds have been overtaken by less bright children from richer ones — and from this point on, the gaps tend to widen still further."

Jerrim and Vignoles wanted to correct the misapprehension that the graph shows that bright working class children in mid-childhood will necessarily fall behind dim middle class children in middle childhood. This misapprehension would be based on the presumptions that: the graph represents a necessary feature of the development of all individuals rather than representing average phenomena; and that it is meaningful and technically possible to identify stable cognitive capabilities at 22 months such that "bright" and "dim" children can meaningfully be identified and classified as such based on tests at 22 months.

Yet, it is not necessary to believe that the groupings are stable, innate or fixed to find the data in figure 2 interesting. The graph shows what happens to the average test scores of different clusters of children in an interaction between an indicator of family origin and average measures of cognitive development, starting very early in childhood. Low social economic status (SES) children in the UK tended (and still tend) on average to fall back relative to middle class children, whatever the early levels of measured ability.

Read (2003) is wrong that the shift between 22 and 42 months was taken by policy makers to be substantive.[1] Much of the chart's role in public debate was as a proxy for a much wider body of research, including more recent analysis of the National Pupil Database and other more recent cohort studies showing how at every stage of education, low income children tend to progress at a slower rate on average than those on higher incomes (Kingdon and Cassen, 2007; Goodman and Gregg, 2010; Magnuson, Waldfogel and Washbrook, 2012). The broad fact is not disputed that the relative access of parents to wealth, income and educational knowledge on average tend to be replicated across generations, in the UK, now and in the past.

It is regrettable that Feinstein (2003) did not include more consideration of the reliabilities of the measures used because differences in reliability at different ages are likely to be responsible for a considerable but unquantified part of the observed pattern of results as children mature. If reliability of measurement increases with age then one might expect the fanning observed in figure 1 and the resulting pattern of figure 2. It is also important that the age ten tests may be more discriminating as tests of cognitive development than the age five tests.

The intention in Feinstein (2003) was explicitly descriptive, aiming to offer a sense of scale of the emergence of the gaps in average scores by children classified in very broad groupings in a very raw and single index of cognitive development.

Measurement error was not dealt with. The aim was to present the actual data, of the kind that is used to test children and award them grades and qualifications, suffering as this does from measurement error, rather than to present corrected trajectories based on modelling assumptions.

The change between ages five and ten years

Jerrim and Vignoles show that under reasonable, though not proven, assumptions the misclassification bias in the average score washes out after the second point of measurement. Therefore, the change in relative position between age five and ten years may be substantive. They note it may be due to a difference in the underlying tests, and suggest therefore this should also be seen as a form of regression to the mean.

The required assumption in their model is that the measurement error at the later ages is not correlated with the measurement error in the earlier scores. The age five tests were taken with a different set of instruments than those at 22 or 42 months, as set out above. They were conducted in the home by health visitors. The age ten tests are a different set of tests again, much more scholastic with a strong focus on maths and reading tested through a longer series of questions asked during a test session in the child's school. Therefore, it seems reasonable to suppose that measurement errors of this sort are not correlated across ages.

However, there are other sorts of possible measurement error that may well be persistent across ages, depending on what is meant by true ability. Some have argued (Gillborn & Youdell, 2000) from a more sociological perspective that low SES children will tend to under-perform in tests of cognitive capability because the tests reflect codes, expectations and structures of power that are themselves class-based.

Perhaps authors in this series might comment on the likelihood and implications of the assumption of zero correlation in measurement error across ages. I certainly agree that the difference in the underlying tests is important, but labelling this regression to the mean in a public debate seems to me to confuse the error resulting from misclassification bias in the early tests with the idea of a genetic basis to social class groups. Although this latter shift could technically be described as "regression to the mean," it is of a very different sort to that of the first kind, and is not adequately explained as necessarily a statistical phenomenon. This is an issue on which further clarification would be useful.

Although concerned with measurement, I see the data in figure 2 as evidence that children from working class families in the 1970 Cohort Study who not only scored well at 22 months on fairly raw tests of cognitive capability, but continued on average to do so at ages 42 months and five years, did not on average translate this ability into school success at age ten at anything like the rate of children in middle and upper class families. The shift from more general features of cognitive development at age five to more scholastic tests of reading and maths is important. Working class children in the 1970s appear to have tended to do worse on average on the age ten scholastic tests than they did on the more general age five tests. Some may argue this is because the age ten tests are better measures of true ability and so better indicate the true abilities of children from the different social class groups. My interpretation is more that the working class children tended to translate their earlier capabilities into success in scholastic test scores less well than did their middle class peers. As has been said many times the graph does not resolve this question.

By age ten it is meaningful and possible to conduct long tests of what children have learned in school. The age five tests are much more generic tests of cognitive capability. So it is informative that whereas middle class children who scored well on the age five copying test tended to score well in later tests, working class children did so to much less of an extent. It may be, as some appear to assume, that working class children who did well on the copying test just got lucky. It seems more likely to me that they just didn't achieve their potential in the later tests. The data do not distinguish between these interpretations.

True ability

There are both statistical and political debates being had and much as statisticians might like the rules of political debate to be reduced to the conventions of statistical debate, this is unlikely to happen. The graph has caused confusion in some quarters because of the difficulty of translating accurate and reasonable interpretations for policy audiences. This has also been difficult for Jerrim and Vignoles whose critique of the false interpretation of the chart has been taken by some as proof that social mobility is inevitable (Saunders, 2011 and Guardian 14 April, 2011 "Government social mobility expert under attack."). In other work Jerrim, Vignoles, Lingam and Friend (2013) show the huge gap between the evidence from structural genetics regarding the heritability of intelligence and that from any biological analysis of actual genetic data in explanation of the social class attainment gap. As discussed further below, it is important in the political debate that the Jerrim and Vignoles model is not taken as proof of its own assumptions, that low SES children are innately less cognitively capable, based on confusion about the meaning of "true ability" in their model. The notion of true ability they use is a statistical convenience, not the suggestion that science or social science has shown in any way that the latent ability gap at each age is in any way innate.

Jerrim and Vignoles base their model on the idea that at every age and moment of development each child has a true level of ability by which they can be ranked on a uni-dimensional scale of cognitive capability, as implied by the first principal component measure used in Feinstein (2003). They apply a standard statistical model in which a true, latent construct is hypothesised to be measured with error, which in this case they define as "true ability" – the specific level of ability of the child with some unspecified degree of stability at the time the measurement was taken. There is a particular definition of true ability at the core of their model,

but it does not concord with a more usual, popular understanding of the notion of true ability, it is a statistical definition.

A second key assumption of their model is that at all ages and moments of development the true component of their variable is socially stratified, that is to say reflective of the degree of wider structural inequality such that the 22 month differences in rank contain and reflect SES differences. They assume that it is a feature of true ability, as well as of test scores. This follows from their implicit definition of true ability as the latent construct at the time of the test, not from presuming that it is a fixed entity, as the Jerrim Vignoles model allows true ability to vary over time. Furthermore, Jerrim and Vignoles do not assume, as does Saunders (2010; 2011) that a social class gap in true ability is a necessary feature of society, occurring necessarily in all social aggregates in all times and places. However, their use of the phrase "true ability" in their statistical modelling does appear to have been taken by some to imply that their model showed that there are stable, biological foundations to the social class attainment gap.

Crucially, Jerrim and Vignoles (2011), add to this the hypothesis that the degree of misclassification bias will vary by SES. Because low SES children are drawn from a group with a lower average score, children drawn from the low SES group who score well early on are more likely to have had, on their terms, over-estimated ability than the similarly scoring high SES children. They go on to say "Low SES children who get defined as high ability have probably had a particularly large random positive error (i.e. a lot of luck) during the initial test."

This is intended to be a statistical observation but we all need to be cautious in how we phrase attempts to explain statistical assumptions by making statements about people.

People, averages and qualitative change

Those I spoke to about the chart understood that it pertains to average rather than individual phenomena and so is an indicator of society and development in general not individual children. That said, I do particularly regret not being much clearer in public use of the study findings that the data in the two charts are averages. They may not describe the trajectories of any individual children. They are representative of a feature of development in general at the social level not of specific individuals.

Jerrim and Vignoles (2013) have shown the error of interpreting figure 2 as showing that at age six bright children from working class families will be overtaken by dim children from upper middle class households in school achievement. Another misuse in the public debate was the elision from average to individual. The Every Child Matters White Paper (HM Government, 2003, p19) stated that "children from a poor background with a high developmental score at 22 months have fallen behind by the age of 10, compared to children from higher socio-economic groups but with a low developmental score at 22 months". This drafting also conflates the average pattern with a universal phenomenon.

The distinction between averages and people is perhaps obvious to readers of this journal but is very easily blurred when statistics are used in wider public discourse. This causes problems for a public debate in which, at the level of society, it is important to know that family assets and capabilities and contexts may tend to impact on school test scores, but in which it would be false to assert that this makes SES the determining factor in the destiny of any specific child. When politicians today claim that figure 1 shows that "the race is over by age five," this is similarly a confusion of individual and average phenomena as it makes a universal of the average, as well as making an exaggerated claim about the average importance of the early years. The data in the graphs above and in all similar analysis tell us something about the average trends, indicating what tends to happen, the tendency in the time and place of the UK 1970 Cohort, not the true history of any individual case.

As Bronfenbrener (1979) and others have shown any framework for intergenerational change involves actors and action at multiple levels of which biological, individual and social levels are particularly important as distinct domains of change. The graph does not begin to address the breadth and complexity of these issues but it does need to be understood in this context if there is to be any discussion of implications for policy. The multi-level approach to understanding longitudinal data set out in Peck, Feinstein and Eccles (2008) emphasises in particular the importance of recognising qualitative change in modelling life course data. The corrections suggested by Jerrim and Vignoles treat cognitive development in the early years as a time-series of a common constructs rather than the emergence of a complex capability

that changes qualitatively through the periods modelled. This is implied by the use of the lines alongside the data points of figure 2, which infuriated many, but it is important not to take this too literally as anything other than the changes in the average scores, that bear a distant relationship to the individual scores and are even more distant from the multi- dimensional and complex development of the individual children. In recognising this, other models might treat these measures in very different ways.

The implication of a literal reading of figure 2 or of the Jerrim and Vignoles correction is that at each age it is unproblematic to compare children in terms of their true ability and stack them up in unique ranks of relative achieved uni-dimensional intelligence. Even at 22 months, their model assumes that the only barrier to achieving this is the technical difficulty of measuring these true ranks. Error results not just from poor measurement but also from the deviation of the "true" distribution of the underlying latent variable from the linearity assumption in the index. So it remains important not to overstate the resulting precision. In their corrections for regression to the mean Crawford, Macmillan and Vignoles (2014) are very careful to label this "high early performance," to distinguish it from anything that might be thought innate, whatever the researchers' intentions. These data in corrected form show a general tendency at a time and a place and between the ages assessed using the specific metrics available, not a fundamental and fixed truth about human beings.

There are a number of different explanations of the facts about cognitive development and social class in the UK. It is conceptually possible that the pattern between 42 months and age ten in figure 2 indicates how capability and context interact to influence outcomes for the children in the 1970 study and hence in general in England, Wales and Scotland in the 1970s. It is also conceptually possible that the pattern is entirely the result of regression to the mean in a very strong sense; that the high scoring working class children were just a group with low true ability with continued luck who eventually got found out as test scores got more accurate. It is true that the data do not discriminate easily between these interpretations. We are left with theory and the wider science to attempt to distinguish them.

Conclusion

The graph shows that children from working class backgrounds in the 1970 cohort with good very early signs of cognitive development were less likely to translate these early signals into good later scores than children from middle class backgrounds. From this graph and many other sources was drawn the line in the strategy: *"Bright children from poorer families tend to fall back relative to more advantaged peers who have not performed as well."*

I wouldn't myself have used the phrase "bright children" but nothing in the Jerrim and Vignoles (2011, 2013) or Read (2003) critiques disprove the statement, as they themselves pointed out (The Guardian 28 April, 2011).

David Willetts MP, at the time Minister for Higher Education in the Department of Business Innovation and Science said subsequently:

> Sometimes over-reliance on one specific piece of evidence can leave you vulnerable. I remember being influenced by Leon Feinstein's very interesting paper for Economica in 2003 called Inequality in the Early Cognitive Development of British Children. He showed that bright poor kids fell behind rich dim kids by the age of 7. I served on Nick Clegg's social mobility group and recommended this powerful evidence to him and he too was impressed and cited it. But Leon's work was challenged by other academics because it was affected by reversion to the mean. The result was that the Guardian ran a piece that the Coalition's social mobility strategy was undermined because the research on which it rested had been disproved. That is not, of course, a reason for giving up on evidence-based policy: but it is a reminder of how careful we have to be in using it.

The question of the age at which supposedly "bright" working class children are overtaken in school performance by supposedly "dim" middle class children is not one that was ever tested or referenced by me. It is regrettable if there was confusion about there being a fixed age of six at which all dim middle class children overtake all

bright working class ones. To be clear the crossover in this form is an artefact of the transparent way figure 2 was constructed and a corollary of figure 1. The point that was important for policy and was referenced in the 2010 Social Mobility Strategy was that throughout childhood in the UK children from low SES homes tend on average to fall back in school achievement relative to children from higher SES backgrounds.

The observed pattern between 22 and 42 months has always been understood by me and those with whom I have discussed the graph as mainly a statistical artefact resulting from measurement error. It has also been, in my experience, well understood that you cannot accurately or meaningfully fix children at 22 months on a scale of absolute and fixed ranks of ability. It would be wrong to define children as "bright" or "dim" on the basis of a set of early tests of cognitive development. Indeed, part of the early interest in the paper was because of the instability it showed in early signals of ability.

In an attempt at explaining the data (Feinstein 2003b, p30) I wrote, "so early scores do matter but so does social class after early childhood. The lesson for policy makers is clear. There is mobility (as one would expect) after 22 or 42 months, but upward mobility is mainly for high or medium SES children. Low SES children do not, on average, overcome the hurdle of lower initial attainment, combined with continued low input. Furthermore, social inequalities appear to dominate the apparent early positive signs of academic ability for most of those low SES children who do well early on."

Some would like to argue this is just an inevitable fact of heredity (Lynn, 2011; Saunders 2011). Some have wanted to claim that these patterns of inequality in development demonstrate underlying genetic continuities such that inequality is inevitable, others that the data show the impact of environment. As I stated in the 2003 paper, the graph cannot answer these questions.

However, there is general agreement that intelligence and school achievement have sufficient fluidity and malleability that only in rare cases is school achievement so fixed that there is no role for education and policy. Heckman (2007) puts it very clearly, based on his model of the production of capability:

> The nature versus nurture distinction, although traditional, is obsolete. Abilities are produced and gene expression is governed by environmental conditions. Behaviours and abilities have both a genetic and an acquired character. Measured abilities are the outcome of environmental influences, including in utero experiences, and also have genetic components.

I think this means it is wrong to interpret this type of longitudinal interaction between early scores and late scores (even if corrected for early reversion to the mean) as the later outcomes of dim or bright children, as though these characteristics were easily discernible in early childhood and fixed.

It is helpful that people are reminded that the graph is not simple and should be considered carefully, bearing particularly in mind the strong classification error between 22 and 42 months. We should remember it was a sample of children from the 1970s.

How children perform in tests matters for many reasons, not least as a signal to themselves and others. How this information is interpreted has a very substantial impact on child achievement and life outcomes (e.g. Dweck, 1986) so in the public debate it is always important to make a clear distinction between the meaning of aggregate statistical data and individual lives.

Subject to issues of modelling and measurement, the pattern of emergence of inequality in development tells us about the nature of inequality at the time and place at which the data are gathered. It is my hope that this debate will lead to further comparative work using diverse methods across diverse datasets to establish what differences are due to measurement, what to modelling and what to time and place.

Acknowledgments

I am grateful to those who have offered comment at various workshops, in particular at seminars at the Centre for the Analysis of Social Exclusion (LSE), at the Institute of Social and Economic Research (University of Essex), at the Genomics Forum (University of Edinburgh) and in a debate with John Jerrim at Portcullis House. Others have provided helpful comment on earlier drafts including Ruth Lupton, Kate O'Neill, Bilal Nasim, Kitty Stewart and Judith Dimant. Remaining errors remain mine alone.

References

Björklund, A., Lindahl, M., & Plug, E., (2006). The Origins of Intergenerational Associations: Lessons from Swedish Adoption Data. *The Quarterly Journal of Economics, 121*(3), 999-1028. http://dx.doi.org/10.1162/qjec.121.3.999

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press

Chamberlain, R. & Davey, A. (1976). Cross-sectional study of Developmental Test Items in Children aged 94-97 Weeks: Report of the British Births Child Study. *Developmental Medicine and Child Neurology*, *18,* 54-70. http://dx.doi.org/10.1111/j.1469-8749.1976.tb03605.x

Crawford, C., Macmillan, L., & Vignoles, A.. (2014). *Progress made by high- attaining children from disadvantaged backgrounds*. Social Mobility and Child Poverty Commission. Department for Education, London

Dweck, C.S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040-1048. http://dx.doi.org/10.1037/0003-066X.41.10.1040

Erikson, R. (1984) Social Class of Men, Women and Families, *Sociology, 18*, 500- 514. http://dx.doi.org/10.1177/0038038584018004003

Feinstein, L., (2003) Inequality in the Early Cognitive Development of British Children in the 1970 Cohort. *Economica, 70*, 277, 73-98. http://dx.doi.org/10.1111/1468-0335.t01-1-00272

Feinstein, L., (2003b). How early can we predict future educational achievement? *CentrePiece 8* (2). Centre for Economic Performance, London School of Economics.

Galton, M. (1886). Anthropological Miscellanea: Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute, 15*, 246–263

Gillborn, D. & D. Youdell (2000). Intelligence, 'ability' and the rationing of education. In Demaine, J. (Ed) *Sociology of Education Today*, London: Palgrave.

Goodman, A. & Greg, P. (Eds) (2010) *Poorer children's educational attainment: how important are attitudes and behaviour?*. York : Joseph Rowntree Foundation.

Heckman, J., (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences, 104*(33), 13250–13255. http://dx.doi.org/10.1073/pnas.0701362104

HM Government (2011). *Opening Doors, Breaking Barriers: A Strategy for Social Mobility*.

HM Government (2003). *Every child matters*. Green Paper, Cm 5860

Jerrim, J., & Vignoles, A. (2011). The use (and misuse) of statistics in understanding social mobility: regression to the mean and the cognitive development of high ability children from disadvantaged homes. *DoQSS Working Paper 11-01*. Institute of Education

Jerrim, J., & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176* (4), 887–906 http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

Jerrim, J., Vignoles, A., Lingam, R., & Friend, A. (2013). The socio-economic gradient in children's reading skills and the role of genetics. *DoQSS Working Paper 13-10*

Kingdon, G. & Cassen, R. (2007). *Understanding low achievement in English schools.* CASEpapers, CASE/118.

Centre for Analysis of Social Exclusion, London School of Economics and Political Science.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley. http://dx.doi.org/10.1002/9781119013563

Lynn, R. (2011). *Dysgenics: Genetic Deterioration in Modern Populations*. Praeger Publishers.

Magnuson, K., Waldfogel, J. & Washbrook, E., (2012). SES Gradients in Skills during the School Years. in: Ermisch, J., Jäntti, M. & Smeeding, T. (Eds) *From Parents to Children: The Intergenerational Transmission of Advantage*, 235-261. New York: Russell Sage Foundation.

Peck, S., Feinstein, L. & Eccles, J., (2008). Pathways through education: Why are some kids not succeeding in school and what helps others beat the odds? *Special issue: Journal of Social Issues, 64,*(1), 1–233. http://dx.doi.org/10.1111/j.1540-4560.2008.00545.x

Piaget, J., (1952). *The origins of intelligence in children.* International Universities Press. http://dx.doi.org/10.1037/11494-000

Read, D. (2003). *Researcher warns that Government Strategy for Social Mobility misled by a statistical trap*. Press release, Warwick University. Retrieved from http://www2.warwick.ac.uk/newsandevents/pressreleases/extreme_statistics/

Saunders, P., (2010). *Social mobility myths*. Civitas, London.

Saunders, P., (2011). *Social mobility delusions.* Civitas, London.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *American Journal of Psychology, 15*, 201-293 http://dx.doi.org/10.2307/1412107

Tu, Y.K., & Law, J. (2010). Re-examining the associations between family backgrounds and children's cognitive developments in early ages. *Early Child Development and Care, 180* (10),1243-12 http://dx.doi.org/10.1080/03004430902981363

Endnotes

[1] http://www2.warwick.ac.uk/newsandevents/pressreleases/extreme_statistics/

[2] https://inequalitiesblog.wordpress.com/2011/06/16/the-rise-and-fall-of-a-killer-chart/

[3] It's a shame journalists so routinely state that the author of a recent study is necessarily the author of all the data. So much work goes into the data that never gets recognised by this approach.

Commentary by

John Jerrim Institute of Education, University College London, UK
j.jerrim@ioe.ac.uk
Anna Vignoles University of Cambridge, UK

Socioeconomic differences in children's test scores: what we do know, what we don't know and what we need to know.

Along with Blanden, Gregg and Machin (2005), Feinstein (2003) is one of the social science papers that has had the biggest impact upon British policymakers since the turn of the Millennium. Despite our subsequent criticism of how these findings have been interpreted (Jerrim & Vignoles, 2013), we encourage readers to remember that Leon Feinstein's 2003 paper offered many important, original and interesting insights. Although blunt in communication, it has always been our intention to engage productively with the important issues raised by Feinstein (2003). With that in mind, we are grateful to both John Bynner (the editor) and Leon Feinstein for their continuing engagement, and the productive platform they have developed for moving the debate forward. Indeed, we hope this Comment and Debate section encourages further work in this area. Our response will be structured as follows. To begin, we will clarify our position on some of the points raised by Feinstein (2015). We then summarise areas where there now seems to be reasonably strong empirical evidence on socioeconomic achievement trajectories and where broad consensus seems to have been reached. This is followed by a discussion of where knowledge is still lacking, or open issues which remain in debate. The conclusion then suggests directions for future research.

Feinstein's contribution to the debate makes the important point that the sample used for his 2003 analysis is a selected subset of the entire cohort, and indeed that there is significant attrition in the British Cohort Study. Undoubtedly this is so and it threatens the external validity of the findings (the extent to which they really are representative of the population as a whole). Further, as attrition is differential, with greater drop out amongst poorer children, the sample of poor students for whom we have complete data may not be representative of the population of poor children. This is an important issue to address when describing the achievement trajectories of poor children. In Jerrim and Vignoles (2013), the Millennium Cohort Study (MCS) data also suffer from sample selection and attrition, albeit at a lower level in the MCS. More work is needed on how sample selection and attrition might impact upon the generalisability of the results. However, it is also important to point out that attrition concerns in and of themselves do not affect the methodological issues of regression to the mean, since in our paper we do a complete case analysis, whereby we examine the trajectories of children in the sample over the entire period. Hence differences between the data sets used by Feinstein (2003) measures and those used by Jerrim and Vignoles (2013) do not explain differences in findings that we attribute to regression to the mean.

Feinstein (2015) also makes the point that the different cognitive achievement tests used across the cohort studies have different scales and meanings, and that the measures used in Feinstein (2003) are ordinal. This is of course correct, and is an issue we consider extensively in Jerrim and Vignoles (2013). Indeed, regression to the mean (RTM) is potentially problematic whether one uses an ordinal scale or not. Nevertheless, it is important that we undertake more work on how best to measure and interpret tests of children's cognitive development at a very early age.

Having clarified these points, we now move on to issues where we believe broad consensus has been reached. First, there are large socioeconomic gaps in children's cognitive skills, and these can be observed from a very early age (indeed, from the very first point measurement of such skills is possible). This is consistent across both Feinstein (2003) and Jerrim and Vignoles (2013), along with a host of other research (e.g. Blanden & Machin, 2007; Crawford, Macmillian & Vignoles, 2014; Cunha, Heckman & Lochner, 2006; Goodman, Sibieta & Washbrook; Jerrim, Vignole, Lingam & Friend, 2014; Jerrim & Choi, 2014; Schoon, 2006). We believe that this represents the main message

that policymakers *should* have taken from Feinstein 2003.

Second, the *absolute* difference in average test scores between socioeconomic groups certainly does not seem to decline as children enter school, and in all likelihood continues to grow. (By the 'absolute' skill gap, we are referring to the actual competencies that high and low SES children display at any given age. It is based upon test scores measured in its original scale – one which has a substantive meaning - and has not been standardised or converted into rank position.) This is due, at least in part, to the increase in the variance of the skills children display as they age. For instance, the difference in what three year olds can do in, say, mathematics is a lot smaller than the variability in mathematics skills displayed by 15 year olds. See Magnuson, Waldfogel and Washbrook (2012) for evidence on this issue from the United States.

Third, although measurement issues cannot be completely ruled out, we believe there is now sufficient empirical evidence to assert that socioeconomic differences in *relative* skills do not appreciably narrow during the school years. (By relative skills, we are essentially referring to the rank order of young people in the achievement distribution). See, for instance, evidence from Feinstein (2003), Magnuson, Waldfogel and Washbrook (2012), Jerrim and Choi (2014), Choi and Jerrim (2015) and Schoon (2006). However, evidence on whether and when the relative skill gap grows ('fans out') or remains stable is (in our opinion) still relatively weak, and susceptible to important (yet little discussed) measurement issues. (For instance, if there is random error/noise in children's test scores, but this decreases as children age, this may also produce the 'fanning out' pattern that is so often cited in this literature). This therefore remains an area where further UK evidence, tackling the important issue of measurement of skill, is needed.

Fourth, with regard to the skill trajectories of initially high (low) achieving children from low (high) SES backgrounds, the striking decline between 22 and 42 months reported in Feinstein (2003: Figure 2) and between 36 and 60 months in Jerrim and Vignoles (2013: Figure 5a[1]) is due, at least in part, to a statistical artefact known as '*regression to the mean (RTM)*'. This should therefore *not* be used by academics or policymakers to stress the importance of the early years, that we are failing 'bright' young people from disadvantaged backgrounds, or to highlight the lack of social mobility in the UK. Rather, the fact that early socioeconomic gaps in achievement are so large is by itself highly suggestive of the importance of the earliest years in more general terms.

Finally, there remains no robust and consistent evidence that initially high achieving young people from poor backgrounds are overtaken by low achieving children from affluent backgrounds in terms of their cognitive skills. Crawford et al. (2014) have attempted to take account of the problem of regression to the mean when measuring the trajectories of initially high achieving students in secondary school and found that high achieving low SES students do decline relative to high SES students between the ages of 11 and 16. However, for this older age group they did not find support for the "crossover" pattern observed by Feinstein (2003). What is also important to remember is that the evidence base does not suggest that the prospects of high attaining (however defined) young people from poor homes are entirely determined by age ten.

If these now represent what we believe to be consensus views, what are the areas of continuing disagreement, and thus where further research is needed? First, although we know socioeconomic differences in cognitive skills emerge early, it is not clear the extent to which this is due to genetics and 'hereditary' factors, and the extent to which this is environmental (or indeed the interaction between the two). Recently, Krapohl and Plomin (2015:3) have argued that 'half of the phenotypic correlation between children's family SES and their educational achievement is mediated genetically' based upon a genome-wide complex trait analysis of 3,000 unrelated children. This is in contrast to some previous research (e.g. Goldberger, 1979, Gould, 2011, Manski, 2011) which either argue against the strength or relevance of such findings. Nevertheless, recent genome-wide association studies (rather than inferred genetic effects based on comparisons across twins) have also indicated that there may be a high degree of heritability in IQ (Davies et al. 2011). Controversial though this is, it might imply that based upon the empirical evidence alone, it is not currently possible to rule out 'hereditary' (a popular explanation by some – e.g. Saunders, 2012) as an explanation for a significant

proportion of the socioeconomic gap in educational test scores. However, it is important to note that the evidence on the importance of hereditary factors is mixed at best.

The picture is undoubtedly complex. Epigenetic studies have suggested that since a child's environment may influence their gene expression, it is by no means straightforward to separate out the effects of hereditary factors and environmental influences, and that the latter influences children's outcomes even in utero (Carey, 2012; Hobcraft, 2012). Further some studies of gene-environment interactions have indicated that whilst differences between socioeconomically advantaged children may be attributable to their genes, environmental factors are more important in socioeconomically deprived environments (Tucker-Drob, Rhemtulla, Paige Harden, Turkheimer & Fask, 2012). Yet the evidence is mixed and partial, with much more research needed. Indeed, on a related note, we must also develop a better understanding of the environmental and genetic mechanisms (and their potential interaction) influencing cognitive development and the growth in absolute socioeconomic skill gaps as children age. Although social scientists typically focus upon the environmental explanations, there is now a growing body of research which suggests that changes in educational attainment over time could be partly due to genetic factors (Haworth, Asbury, Dale & Plomin, 2011).). Rather than shy away from this issue, social scientists should engage more with geneticists and their data – developing a better understanding of how genes may influence cognitive skill growth (including via potential interactions with the environment).

We also still know very little about the educational progress made by initially high-achieving children from disadvantaged backgrounds. Both Feinstein (2003) and Jerrim and Vignoles (2013) have methodological limitations, with the trajectories in both papers subject to a high degree of uncertainty. (As just one example of uncertainty, neither paper presents confidence intervals. But sampling variation is likely to be large, given the small sample sizes of the high/low achieving groups in the data those studies use). Although Crawford, Macmillian and Vignoles (2014) have recently added to the evidence base, further work, using better data and more sophisticated methodology to overcome the RTM problem, is clearly still required.

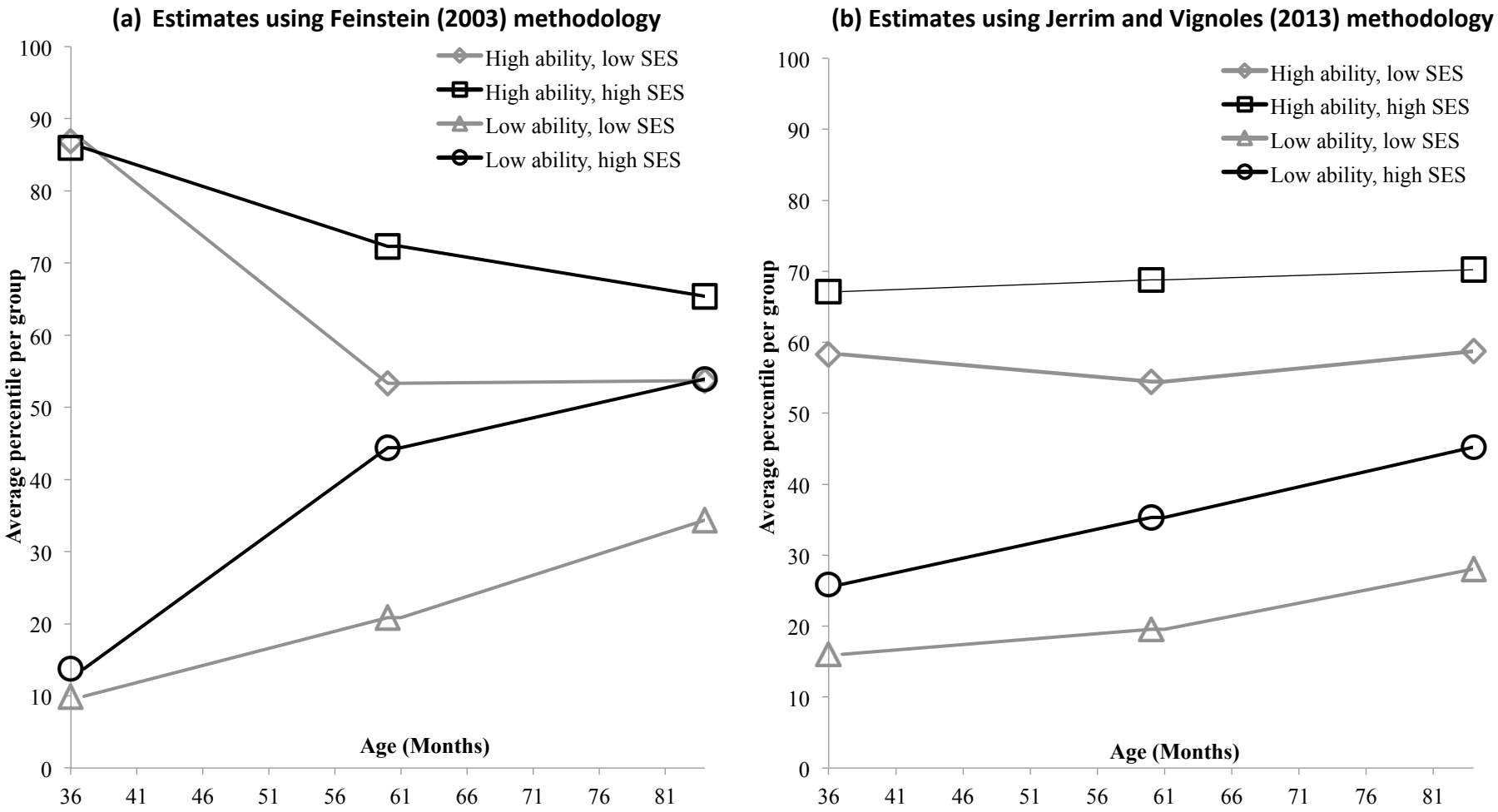
Finally, regarding mean differences in test scores by SES group (not stratified by initial achievement), further detail is needed on the descriptive patterns observed. For instance, if the socioeconomic gap in children's test scores really does increase as children age, is this being driven by the poorest children in society falling behind the rest of the population? Or is it because the most affluent 20% are pulling away from everyone else? These two scenarios would likely warrant quite different policy responses. Our reading of the literature suggests that it is less likely due to the former and more likely to be attributable to the latter (see Goodman & Gregg, 2010; Jerrim & Vignoles, 2015), though again further work is needed.

What then do we suggest are the most pressing issues for those looking to move forward the debate? To begin, more and better data is needed, with greater consideration given to the measurement properties of children's test scores. Despite their many advantages, the UK's 1958, 1970 and 2000 cohort data have some limitations in this respect. Comparable tests have not been conducted at more than two time points for example, with the survey documentation lacking sufficient discussion on possible measurement error and test reliability. The new forthcoming cohort (Life Study) offers potential opportunities to improve the evidence base. It is essential that test scale measurement and psychometric properties are key in any future cohort collecting data on early-life cognition. With regard to existing data sources, the survey organisers and funders should attempt to retrospectively consider such issues. In general, the properties of the achievement data contained within the cohorts need to be more thoroughly investigated and better understood.

Next, more sophisticated methods need to be developed and applied in this area. These should ideally be able to account for the possibility that there is 'systematic' error in the test-score data, including potential measurement bias. In other words, future work needs to move beyond the simplistic assumption made in Jerrim and Vignoles (2013) that any error in test scores is simply random noise. Allowing for systematic error is likely to enable us to get closer to the 'truth'. Although we note the potential for latent variable methods such as growth curve modelling to do this (as per Von

Stumm & Plomin, 2015), this should be accompanied by a clear explanation as to how it overcomes the RTM problem and a simulation study demonstrating the conditions under which it works (and the assumptions being made). After all, the beauty of Feinstein (2003: Figure 2) was its simplicity. (This was one of the key reasons why we proposed the simple adjustment in Jerrim and Vignoles 2013 – reproduced in figure 1 - to maintain this simplicity). Future work should continue with such clarity of presentation to keep policymakers engaged in this matter – while, of course, ensuring the most robust and convincing methods are applied.

Figure 1. Estimated cognitive gradients in MCS when using different methodologies



Note: Reproduced from Jerrim and Vignoles (2013:Figure 5). Estimated cognitive trajectories based upon the MCS. The left hand panel refers to estimates using methodology of Feinstein (2003). The right hand panel is the equivalent figures when applying the methodology proposed by Jerrim and Vignoles (2013

Third, better use needs to be made of existing resources to tackle the issues we have raised. The Twins Early Development Study (TEDS) is a prime example, which contains detailed information on both children's genetic and parental investments for a large sample of UK twins, who have been tested at multiple points throughout childhood (from age two through to age 18). Such data has the potential to provide new descriptive information on SES trajectories. Once the cognitive trajectories are firmly established, this will be the next vital step in this line of research. TEDS is an underutilised dataset by social scientists, and one we believe can potentially address many of the issues described in this paper. We therefore strongly encourage any social scientist looking to conduct further work in this area to consider seeking to use this MRC funded dataset.

Finally, we end on a note of caution. Although there is a desire amongst policymakers to know how policy should respond to counter the deleterious effects of SES on achievement, it is important that we walk before we run. We first need to be certain of the descriptive trajectories regarding how cognitive skills develop differentially across socioeconomic groups. Here, high quality data and robust methodologies are key. It is only once this first stage is complete, and to a satisfactory standard, that we should then attempt to disentangle cause from effect and develop the appropriate policy response.

References

Blanden, J., Gregg, P. & Machin, S. 2005. *Intergenerational mobility in Europe and North America*. The Sutton Trust report. Retreived from http://www.intouniversity.org/sites/all/files/userfiles/files/Sutton%20Trust%20Social%20Mobiliity.pdf

Blanden, J. & Machin, S. (2007). *Recent changes in intergenerational mobility in Britain*. Sutton Trust, London. Retreived from http://www.suttontrust.com/public/documents/summaryintergenerationalmobility.pdf .

Carey, N. (2012). *The Epigenetics Revolution: How Modern Biology is Rewriting Our Understanding of Genetics, Disease, and Inheritance.* New York: Columbia University Press.

Choi, A. & Jerrim, J. (2015). The use (and misuse) of PISA in guiding policy reform: the case of Spain. *IEB working paper series*. Retreived from http://www.ieb.ub.edu/phocadownload/documentostrabajo/doc2015-6.pdf http://dx.doi.org/10.2139/ssrn.2580141

Crawford, C., Macmillan, L., & Vignoles, A. (2014). *Progress made by high attaining children from disadvantaged Backgrounds*. Research report.

Social Mobility and Child Poverty Commission. Department for Education, London

Cunha, F. Heckman, J. & Lochner, L. (2006). Interpreting the Evidence on Life Cycle Skill Formation. In E. Hanushek & F. Welch *(Eds)Handbook of the Economics of Education*. Pp. 697-812. Amsterdam: Holland North.

Davies, G. Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D. & Deary, I.J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *MoI Psychiatry 16*(10), 996-1005. http://dx.doi.org/10.1038/mp.2011.85

Feinstein, L. (2003). Inequality in the early cognitive development of British Children in the 1970 cohort. *Economica 70*, 73-97. http://dx.doi.org/10.1111/1468-0335.t01-1-00272

Feinstein, L. (2015). Social class differences in early cognitive development and regression to the mean. *Longitudinal and Life Course Studies 6*(3)

Goldberger, A. (1979). Heritability. *Economica 46*(184), 327-347. http://dx.doi.org/10.2307/2553675

Goodman A., Sibieta L., & Washbook E. (2009). Inequalities in educational outcomes among children aged 3 to 16. Final report for the National Equality Panel, Institute for Fiscal Studies, London. Retrieved from http://sta.geo.useconnect.co.uk/pdf/Inequalities%20in%20education%20outcomes%20among%20children.pdf.)

Goodman, A. & Gregg, P. (2010). Poorer children's educational attainment: how important are attitudes and behaviour? Report for the Joseph Rowntree Foundation. Retrieved from http://www.jrf.org.uk/system/files/poorer-children-education-full.pdf

Gould, S. *The Measurement of Man* (second edition), W. W. Norton & Company

Haworth, C., Asbury, K., Dale, K. & Plomin, R. (2011). Added value measures in education show genetic as well as environmental influence. *PLoS One 6*(2), e16006. http://dx.doi.org/10.1371/journal.pone.0016006

Hobcraft, J. (2012). Epigenetics and the social sciences: progress, prospects and pitfalls, paper presented at the ESRC International Grand Challenge Symposium, 25-26 June 2012. Retrieved from http://www.esrc.ac.uk/_images/Epigenetics-symposium-background-paper_tcm8-21507.pdf.

Jerrim, J. & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society Series A 176*(4), 887 – 906. http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

Jerrim, J., Vignoles, A., Lingam, R. & Friend, A. (2014). The socio-economic gradient in children's reading skills and the role of genetics. *British Education Research Journal* DOI: 10.1002/berj.3143 http://dx.doi.org/10.1002/berj.3143

Jerrim, J. & Choi, A. (2014). The mathematics skills of school children: how does the UK compare to the high performing East Asian nations? *Journal of Education Policy 29*(3), 349-76. http://dx.doi.org/10.1080/02680939.2013.831950

Jerrim, J. & Vignoles, A. (2015). University access for disadvantaged children: A comparison across English speaking countries. *Higher Education 1007*/s10734-015-9878-6

Krapohl, E. & Plomin, R. (2015). Genetic link between family socioeconomic status and children's educational achievement estimated from genome-wide SNPs. *Molecular psychiatry*. http://dx.doi.org/10.1038/mp.2015.2

Magnuson, K., Waldfogel, J. & Washbrook, E. (2012).The development of SES gradients in skills during the school years: evidence from the U.S. and U.K. In Ermisch, J., Jantti, M., & Smeeding, T. (Eds) *Inequality from Childhood to Adulthood: A Cross-National Perspective on the Transmission of Advantage*, New York: Russell Sage Foundation.

Manski, C. (2011). Genes, eyeglasses, and social policy. *Journal of Economic Perspectives 25*(4), 83-94. http://dx.doi.org/10.1257/jep.25.4.83

Saunders, P. (2012). *Social Mobility Delusions.* London: Civitas.

Schoon, I. (2006) *Risk and Resilience: Adaptations in Changing Times,* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511490132

Tucker-Drob, E., Rhemtulla, M., Paige Harden, K., Turkheimer, E. & Fask, D. (2012) Emergence of a gene × socioeconomic status interaction on infant mental ability between 10 months and 2 years, *Psychological Science 83*(2), 743-757.

Von Stumm, S. & Plomin, R. (2015). Socioeconomic status and the growth of intelligence from infancy through adolescence. *Intelligence 48*(1), 30-36. http://dx.doi.org/10.1016/j.intell.2014.10.002

Endnotes

[1]This graph is reproduced in figure 1 below.

Commentary by **Harvey Goldstein** University of Bristol, UK
h.goldstein@bristol.ac.uk
Robert French University of Bristol, UK

Differential educational progress and measurement error

Abstract

The debate around Feinstein's original (2003) analysis is crucially dependent on technical issues associated with definitions of measurement error and how the presence of such error can be adjusted for. In this commentary we explore the ways in which measurement error can be incorporated within regression (and other) models to obtain valid inferences. We suggest that there are flaws in both Feinstein's original analysis and his current response to criticism, as well as problems with some of the existing technical critiques. We conclude that there is reasonable evidence for increasingly divergent educational achievement among social groups as children move through schooling.

Keywords: Social class differences, measurement error, regression to the mean, educational achievement.

Introduction

The publication of Feinstein's original analysis (2003) generated both substantive and methodological concerns with import for social policy. In the latest response to criticisms of his earlier work Feinstein, (2015) broadly defends his earlier findings while at the same time appearing to accept that there are legitimate concerns about the statistical model underpinning his analysis. The main purpose of our paper is to clarify the underlying statistical issues, and in the process of doing this we will comment on Feinstein's conclusions.

We reanalyse the dataset Jerrim and Vignoles (2015) used to critique Feinstein (2003), since this will allow us to illustrate the key technical issues. We then discuss the relevance of our model estimates to the questions originally raised by Feinstein. Before doing this, in the interest of clarity, we need to comment on some of the assertions made by Feinstein (2015).

He claims in section two that his analysis "was explicitly descriptive" as opposed to constituting a statistical model. He says "The aim was to present the actual data…, rather than to present corrected trajectories based upon modelling assumptions". We argue that presenting data, whether based upon a sophisticated statistical model or a simple statistical model such as that which lies behind figure 1 in Feinstein's (2003) paper , is intended to convey an inference about the underlying social process that is generating the data. Feinstein is not presenting 'actual' data, rather he is presenting a summary based upon a particular manipulation (i.e. modelling) of the data with the intention to convey something about how social class differences in achievement are changing over time.

The second issue is Feinstein's use of the terms 'regression to the mean' and 'true scores'. Regression to the mean, which is also used by Jerrim and Vignoles (2012), as introduced by Francis Galton simply occurs when the correlation between two measurements over time is less than one, as is the case with heights of fathers and sons. The notion of measurement error is entirely separate, although if one had perfectly correlated 'true' measures then the addition of random measurement error to these would lead to the same mathematical result. Using standardised measures, so that each measurement is on the same scale with a mean of zero and standard deviation of 1, the mean second occasion score for those with a given high score, x, on the first occasion will be smaller than x. This raises the issue of what is meant by 'true score'. Feinstein does have a useful discussion on this, but is not entirely clear about the 'statistical' notion of true score. This is a conceptual notion that proceeds from the common observation that the actual score that a child obtains on a test will depend on the actual

items chosen for the test plus other factors that might be considered 'transient' such as time of day, test environment etc. Most test constructors (although sadly it appears not the providers of the tests in question) provide estimates of this 'unreliability' or measurement error so that it can be taken account of by data analysts.

The statistical model

We start by presenting the original plot from Jerrim and Vignoles (2012) in figure 1. It illustrates the approach adopted by them and Feinstein (2003). The data consist of four measurements on a sample of children from the 1970 British Births cohort study at the occasions of 22, 42, 60 and 120 months. In this note we will use only two occasions in order to illustrate our points.

Figure 1.

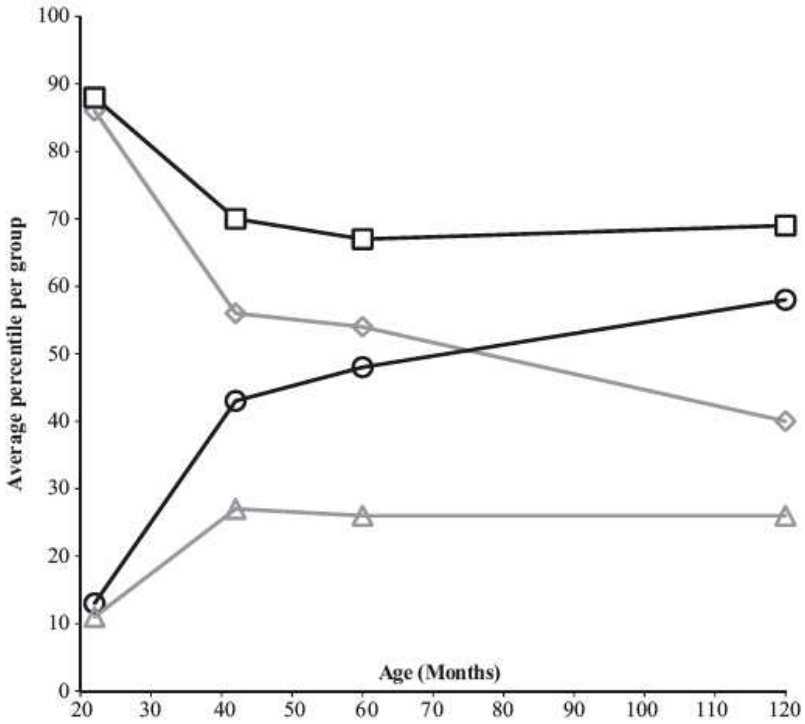


Fig. 1. Development of high and low ability children by socio-economic group—evidence from the existing literature (adapted from Feinstein (2003); based on 1970 British Cohort Study data): ◇, high ability–low SES; □, high ability–high SES; ○, low ability–high SES; △, low ability–low SES

We also use the more extensive Millennium Cohort Study (MCS) dataset, as do Jerrim and Vignoles, in our analysis at ages of approximately three and five years. The number of children available for our analysis, after excluding the small number of cases with missing data (effectively missing at random) is 10,071.

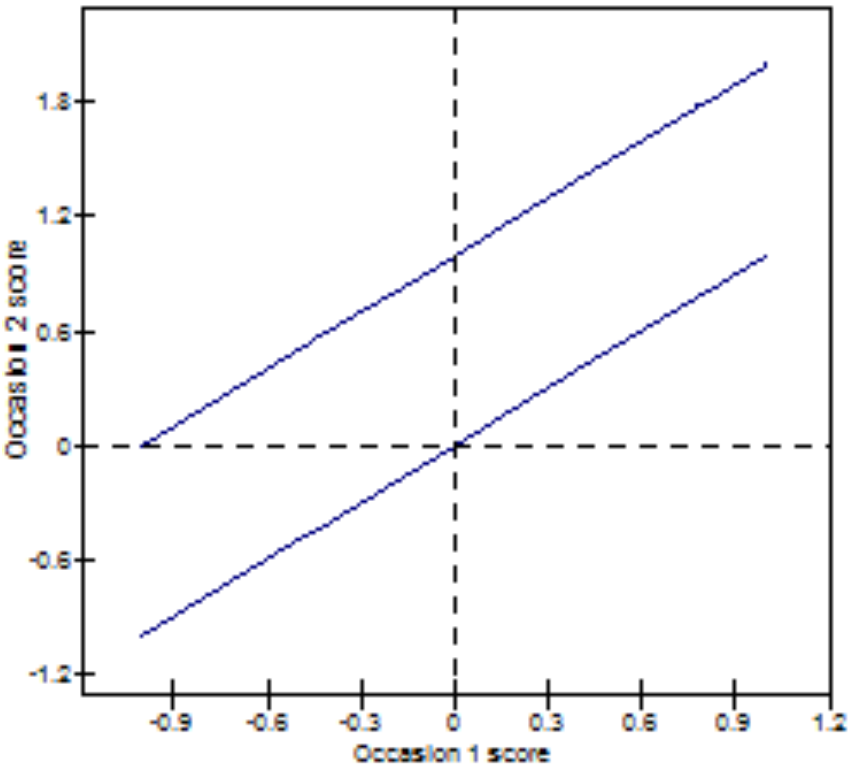
As is clear from figure 1, the previous analyses use the first occasion ability measure by forming two ability groups, the bottom decile and the top decile. Feinstein uses the actual ability measure itself to form these groups whereas Jerrim and Vignoles use a highly correlated 'surrogate' measure. This is because their method requires measurement errors in the grouping variable to be uncorrelated with those in the ability measure itself and they are prepared to make the assumption that this is satisfied by this surrogate measure (see below). In our exposition this is unnecessary, as we shall show. Socioe-conomic status (SES) groups are defined by Jerrim and Vignoles using an income measure with lower and upper quartile thresholds

so that about 25% are classified as high SES and 25% low SES. They obtain essentially the same results as those given in figure 1, where the inference is that for children of the same ability at the first occasion those in the high SES group show greater progress than those in the lower SES group, and indeed that the initial high ability low SES children start doing worse than the initial low ability but high SES children by just after the third occasion.

At this point it is worth remarking that some care has to be taken with this form of adjustment for initial achievement, and this is also pointed out by Jerrim and Vignoles. Since the low SES children at the first occasion on average have a markedly lower ability – about 0.5 standard deviations below the mean and high SES children have a mean ability about 0.3 standard deviations above the mean, the mean abilities for the SES groups may also be expected to differ when we study just the low ability or high ability children. In fact, for the MCS data the mean abilities are about the same for the high and low SES groups within the group of high ability children, but for the low ability children the low SES children have an average ability about 0.3 standard deviations below the average ability for the high SES children. This result, of course, is based upon the observed scores but in general, for unimodal distributions, we would expect a similar result for true scores, that is with no measurement errors. We discuss the definition of true score and measurement error in equations (2) & (3) below. In such a case, with known true scores and an overall average difference between upper and lower quartile SES groups of 0.8, the average difference between high and low SES groups is about 0.2 for low ability children and also 0.2 for high ability children. The more extreme the ability groupings that are used the less this difference will be. Thus in general, even if there were no measurement errors, and no changes in ability over time, we would expect the low SES low ability children on average to have lower average ability over time than the low ability high SES children, and the same principle would apply to the high ability children, simply because they initially have different average abilities because of the way they were selected. In other words, grouping on initial ability in this fashion, whether using a surrogate for true score, or just the observed score, is not generally an adequate adjustment for initial ability, even when we are dealing with true scores without measurement error. This suggests that this kind of approach to studying changing social gaps over time is inherently flawed and is not to be recommended.

To adjust properly for initial ability differences, in order to study progress over time, we will adopt a modification of what is a commonly used approach whereby we model the second occasion ability score as a suitable function of the first occasion score, together with SES and possible interactions. Figure 2 shows a hypothetical example of the results of such a model where for any given true occasion one score the high SES group has a higher than expected true score at occasion two.

Figure 2. Hypothetical relationship for two SES groups



The parallel lines in this graph imply that the mean group difference is the same for all individuals at each occasion one score and we note that this adjusts for the initial ability without any ability grouping involved. This approach is able to show the SES differences for the full dataset rather than just two extreme groups. Interest lies in whether in reality the slopes of the lines in fact differ or whether the relationship may be non-linear. In practice, of course, where we have 'observed' scores that include measurement error, rather than 'true' test scores we will need to adjust for this, which may well change the inferences we make. In the following analysis we shall look at the effects of SES on progress before and after adjusting for measurement error. We do not here present details of the measurement error adjustment, since these form the basis of another paper in preparation. A basic reference, however, is Richardson and Gilks (1993) who outline a Bayesian modelling approach of which ours is a further extension, most notably by allowing interaction terms that include measurement errors. The presence of interaction (and power) terms as in model (1) below is needed in our analysis. In fact, consistent procedures for standard regression models that allow adjustment for measurement error have been known about for over 40 years and Goldstein (1979) uses these in an analysis of 1958 cohort data to study precisely the question of differential progress for different social groups.

A model for measuring differential group progress

Our basic model can be written as

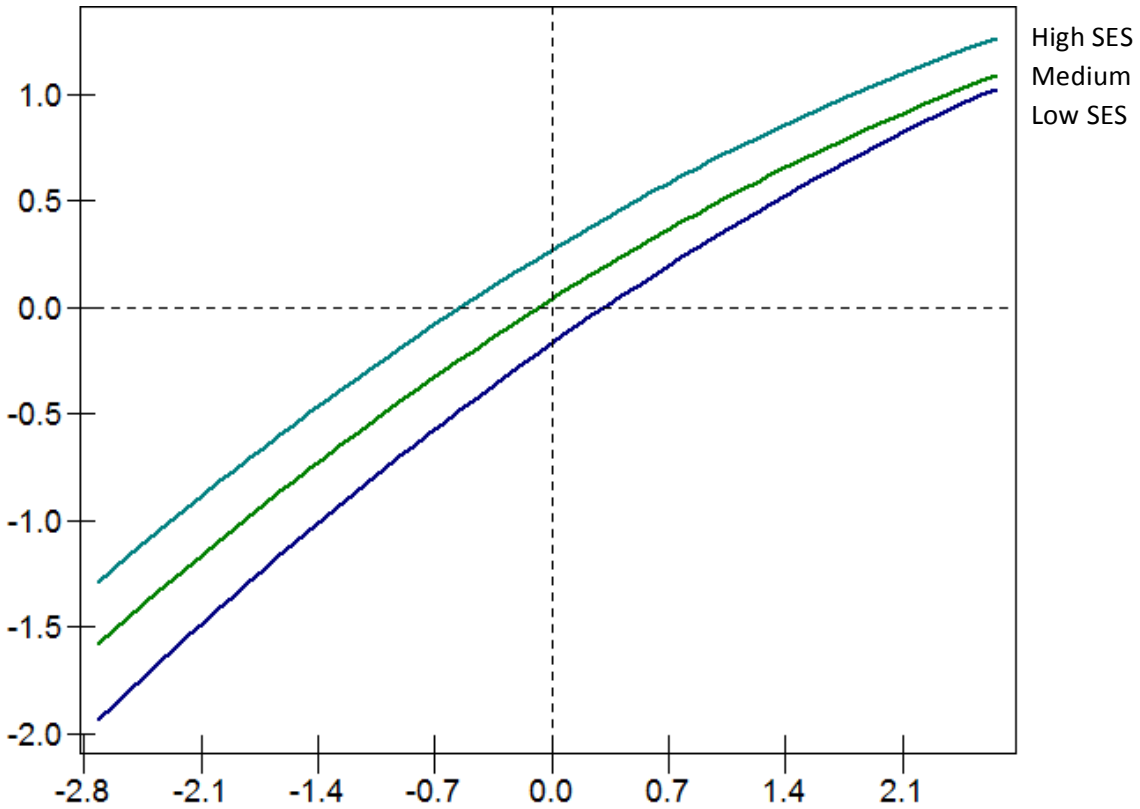
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_{2i} z_{1i} + \beta_{3i} z_{2i} + \beta_{4i} x_{1i}^2 + \beta_5 x_{1i} z_{1i} + \beta_{6i} x_{1i} z_{2i} + e_i \quad (1)$$

Where x_{1i} is the occasion one ability measure, in this case a reading (vocabulary) score, and for which we have used a quadratic term also to reflect the nonlinear nature of the relationship as seen in figure 3. The terms z_{1i}, z_{2i} are two dummy

variables denoting the lower and upper SES quartile groupings with the middle group as the base category. In this case SES groupings are based upon a measure of income, for further details of the variables used see Jerrim and Vignoles (2012). Note that we have also allowed for a possible interaction between initial ability and SES group. We have chosen the SES thresholds such that they contain 26%, 48% and 26% of the sample for the low, medium and high groups respectively.

Column 1 (σ_m^2=0) of table 1 and figure 3 show the results from fitting this model without any adjustment for measurement error, σ_m^2 being the measurement error variance as explained in (3) below.

Figure 3. Predicted age five test score by age three test score for three SES groups. Standardised scores.



We see that for every age three score those in the highest SES group are predicted to have the highest age five score, followed by the middle SES group followed by the lowest SES group. This is consistent with Feinstein's argument for an increasing gap emerging between SES groups. Note also the interaction indicating that the greatest gap between progress made by different groups is for the initial low achievers. In effect these results are similar to those of Feinstein. Thus a low achiever at approximately the lowest quartile position in the low SES group has an expected age five score about 0.7 below a high SES low achiever, whereas a low SES high achiever at approximately the highest quartile position has an expected score about 0.3 below the corresponding high SES child. A high SES child at about the lowest quartile position on age three ability has a predicted age five score of about -0.5 which is what would be predicted for a low SES child a little below the mean.

We now extend the approach used above to adjust for measurement error. The actual size of the measurement error, its variance, is unknown. Jerrim and Vignoles in their simulation use a value of 0.25, equivalent to a reliability of 0.75 for these data. The reliability is defined as the variance of the true scores divided by the variance of the observed scores, where the true score variance is simply the observed score variance minus the measurement error variance. For simplicity we assume that the measurement error variance in the predictor and the response, the reading score at five years, are the same. Now, the observed correlation between the three and five year scores is 0.55 so that this provides the lower limit for the reliability, *R,* since the adjusted true score correlation is estimated by $0.55/R \leq 1$. In fact the correlation varies from 0.56 in the low SES group to 0.44 in the high SES group, and the measurement error variances may well differ also, but we do not explore this further. For the reliability values 0.75 and 0.65 the estimated true score correlations are 0.73 and 0.85 and the latter would seem to be a reasonable upper estimate in practice. We will carry out analyses for these two values. We shall explore the effects of using these reliabilities to judge the sensitivity of the results to the value chosen.

Model (1) is thus augmented by adding the standard measurement error assumption

$$x_{1i} = X_{1i} + m_i, \quad \sigma_m^2 = 0.25 \qquad (2)$$

and

$$y_i = Y_i + q_i, \quad \sigma_q^2 = 0.25 \qquad (3)$$

Where in (2) X_{1i} is the 'true score' and m_i is the measurement error assumed to have a normal distribution with zero mean. Likewise for the response *Y* in (3). Additionally we assume that the measurement errors at occasion one and two are independent, since these are well separated in time. The test scores are all standardised to have zero means and standard deviations of 1.

In addition to the uncertainty surrounding the term 'true score' a further complication is that the size of any measurement error may vary according to individual characteristics. Thus, for example it may be higher for some SES groups than others. We shall not pursue such matters here, and assume that we have a common measurement error variance.

Table 1 shows the results from fitting models (1) + (2), using Markov Chain Monte Carlo estimation with default priors (see Richardson and Gilks, 1993, for further details).

Table 1. Age five reading score related to age three reading score and SES. Different amounts of measurement error variance, σ_m^2 and reliability

Parameter	σ_m^2=0, $R = 1.0$	$\sigma_m^2 = 0.25, R = 0.75$	σ_m^2=0.35, $R = 0.65$
β_0	0.038 (0.013)	0.052 (0.013)	0.042 (0.014)
β_1	0.494 (0.012)	0.712 (0.016)	0.874 (0.019)
β_2	-0.207 (0.021)	-0.092 (0.022)	-0.006 (0.025)
β_3	0.232 (0.021)	0.168 (0.022)	0.125 (0.024)
β_4	-0.037 (0.006)	-0.090 (0.011)	-0.112 (0.014)
β_5	0.054 (0.021)	0.030 (0.029)	0.024 (0.032)
β_6	-0.022 (0.022)	0.0016 (0.0300)	0.021 (0.029)
σ_e^2	0.668 (0.009)	0.515 (0.009)	0.431 (0.011)

Estimation by MCMC: burn in = 500, iterations = 1000.

We see that for the upper and lower SES groups, with a reliability of 0.75 the mean difference $(\beta_3 - \beta_2)$, that is for students with the mean occasion one ability of zero, is reduced from 0.44 to 0.26 SD units and this changes little across the age three reading score since the interaction terms are small, and in fact not significant at the 5% level. When the reliability drops to 0.65 the SES difference becomes 0.13.

Finally, we make a brief comment on the method used by Jerrim and Vignoles where they use an instrumental variable approach. The instrument they use is an 'auxiliary' or 'instrumental variable' taken at the same time as the test of interest at age three and they assume that this is uncorrelated with measurement error in the test of interest. This does seem to us a very strong assumption, especially since the tests were taken on the same day. Furthermore, for instrumental variable methods where it is likely that the instrument is uncorrelated with measurement errors in the test of interest, the instrument itself will tend not to be a very good predictor of the true score. This implies that it will often tend to lack statistical power. In addition there is a substantive problem in that using the instrument as a measure of 'ability' assumes that it is the same measure of 'ability' that is being measured by the test of interest, in other words it is what is known as a parallel test. This does, however, seem questionable, and Feinstein (2015) also picks up on this point. For these reasons we have adopted the above approach, but we do need to emphasise that a sensitivity analysis, using more than one estimate for the measurement error variance is important.

Concluding remarks

The debate about how to study differential progress of children from different socio-economic backgrounds is clearly important and has illustrated the crucial nature of the modelling assumptions that need to be made. Much depends upon knowledge of the quality of the measures used to define educational or other performance, and it is just such information that is typically absent. In particular the reliability of the tests used needs to be available, or at least a plausible range for such values. Such information should ideally be provided by the constructors and suppliers of the tests and users of the data need to be provided with such information, or at the very least made aware of the need for it. The contribution by Jerrim and Vignoles is therefore important in raising this issue, and in this short note we have suggested a comprehensive approach to studying the issue, using a method that allows quite general models, including multilevel ones, to be fitted. Even without such an extension, however, simple moment based estimators using observed variances and covariances of the observed variables corrected for reliability, are available that will generally provide insight into the extent to which inferences are changed when measurement errors are allowed for. Goldstein (1979) describes this approach in the analysis of data from the 1958 British birth cohort in a similar analysis of differential progress. He showed that reasonable amounts of measurement error, when adjusted for, reduced the size of the estimates for differential social class progress and in the case of change from 11 to 16 years in Mathematics reduced the unadjusted difference to a negligible amount. Nevertheless for change between seven and 11 years for both reading and mathematics attainment, there were still differences after adjustment. Unfortunately this evidence was ignored by Feinstein, as was any procedure for handling measurement error.

As far as the substantive issue goes, our exploration of the data shows support for the proposition that there is indeed a widening performance gap especially for children of high SES parents compared to the remainder. Thus, we do not concur with the claim by Jerrim and Vignoles that there is no convincing evidence to support this. To this extent Feinstein's original conclusions are broadly supported, but the actual extent of the widening gap is still an open question, although likely to be much less and more nuanced than he has claimed. Nevertheless, even with our low estimate of reliability (0.65) we still estimate that those from the high SES group are on average 0.13 of a standard deviation ahead of the remainder at age five given the same achievement at age three.

To be fair, Feinstein is far from alone in ignoring these methodological issues. In a recent highly quoted report looking at the progress between 11 and 16 years of 'bright but disadvantaged students', Sammons, Toth and Sylva (2015) also fail to recognise the problems associated with conditioning on a high achieving group, in their case the top third of students, and they also fail to take account of measurement error.

The debate engendered by Feinstein's original paper and various critiques, especially that of Jerrim and Vignoles, has clearly been a difficult one for policymakers, turning as it does on a rather poorly understood set of technicalities. In our view all of this suggests that a more cautious, long term attitude should be taken towards such research findings. Social research is a highly contested area, whether published in a 'reputable' journal or as a non peer-reviewed report to a sponsor.

Policymakers would do well to promote a wide debate about any findings that appear important, where technical and interpretational issues are debated in terms that are widely accessible, and where other relevant research can be discussed. This would be in everyone's interests, not least that of the policymakers themselves who would more often be seen as interested in pursuing useful knowledge rather than advancing their own predilections.

References

Feinstein, L. (2003). Inequality in the early cognitive development of British children in the 1970 cohort. *Economica, 70*, 73-97. http://dx.doi.org/10.1111/1468-0335.t01-1-00272

Goldstein H. (1979). Some Models for Analysing Longitudinal Data on Educational Attainment (with discussions). *Journal of the Royal Statistical Society Series A 142* (4) 407-442 http://dx.doi.org/10.2307/2982551

Jerrim, J. & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society Series A 176*(4), 887 – 906. http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

Richardson,S. & Gilks,W. (1993). Conditional independence models for Epidemiological studies with covariate Measurement error. *Statistics in medicine, vol. 12,* 1703-1722. http://dx.doi.org/10.1002/sim.4780121806

Sammons, P., Toth, K. & Sylva, K. (2015). *Subject to background: What promotes better achievement for bright but disadvantaged students?* Oxford University Department of Education. Retrieved from http://www.suttontrust.com/researcharchive/subject-to-background/

Commentary by **Elizabeth Washbrook** University of Bristol, UK
Liz.Washbrook@bristol.ac.uk
RaeHyuck Lee Columbia University, US

Beyond the Feinstein chart: Investigating differential achievement trajectories in a US cohort

Introduction

The policy and media attention attracted by Feinstein's (2003) paper revolved chiefly around a single "killer chart", depicting trajectories of the average test scores over the course of childhood for four distinct groups of children – "high-" and "low-" scoring children at 22 months of age from high- and low-socioeconomic status (SES) backgrounds respectively. As will be the case in any killer chart, the compelling telling of a visual story can only be achieved by a great deal of selection and simplification of the underlying messy data. This is of course precisely what the analyst sets out to do – to understand and communicate the relationships that are meaningful in a substantive sense. Feinstein's chart was a hugely valuable contribution in the way that it captured imaginations and drew attention to the topic of the developmental trajectories of children from disadvantaged backgrounds. But having kick-started the debate, it is unfortunate that certain features of the chart seem now to be taken as inherent features of the problem at hand, rather than particular choices made for their convenience or visual impact. In this paper we echo Feinstein's (2015) call for more developed analyses of SES differentials in children's trajectories: analyses that build on his original chart but that consider the full range of children's achievement levels, draw on more sophisticated statistical methods, and that take particular care in the interpretations placed on the data. The ideas are illustrated briefly with a longitudinal analysis of the reading skills of a recent cohort of children from the United States.

Our starting point is that the real question of interest that underlies Feinstein's chart is whether trajectories of development tend to diverge, on average, between children who begin with an identical level of baseline achievement but come from different social backgrounds. This is essentially a question about the *timing* of the evolution of SES gaps, about how far inequality observed at the end-point is "locked in" at the time of baseline assessment versus how far it develops subsequently. Jerrim and Vignoles (2013) highlight one important statistical reason why Feinstein's chart might give inaccurate answers to this question: measurement error in baseline test scores will tend to exaggerate the extent to which low SES children fall behind higher SES children who begin in the same initial achievement group. Although this point is undoubtedly valid, it is unfortunate that Jerrim and Vignoles' critique has been interpreted as evidence that the relationships depicted in Feinstein's chart are entirely spurious, as their analysis in no way proves this to be the case. In fact, another feature of Feinstein's analysis – the use of cut-points to classify children into discrete "low" or "high" achieving groups at baseline – also hinders our ability to define identical starting points for children from different backgrounds. There are statistical problems with this approach, plus it may have obscured the fact that the issue considered by Jerrim and Vignoles is in fact a classic measurement error problem, for which a highly sophisticated toolkit of methods for correction and assessment of sensitivity has long been available.

Technical issues aside, we also welcome the opportunity to comment on some of the interpretations that have been applied to Feinstein's findings in general. We question the legitimacy and possible implications of labelling a child as being of high or low ability on the basis of a narrow set of tests at a particular age, and also the disproportionate focus that some have placed on the outcomes of "high ability" low SES children as compared to other low SES children. We begin with this discussion of the messages that have been taken from Feinstein's work in the world of policy and practice. We then go on to address a number of statistical issues in the way that the chart, and some related research, has been constructed, before presenting our analysis of trajectories of reading achievement among US children born in the early 1990s. We show that low SES children systematically failed to keep pace with higher SES

children who began with identical reading skills in kindergarten, and that these inequalities cannot be explained by errors in test score measurement. We then offer some concluding remarks.

Issues of interpretation

The question of whether it is either meaningful or ethical to label some young children and not others as "highly able" on the basis a test taken at one particular age is one discussed in detail by Feinstein (2015), and we strongly welcome the problematizing of this issue. We believe that the uncritical use of this term, as well as of the even more emotive adjectives "clever", "bright" and "dim" in this context is unscientific on the part of academics, and politically loaded on the part of any commentator. We agree that, when characterising the starting point of whatever trajectories are being measured, terms like "initially high performing" or, as we use here, "initially high achieving" should be used as much as possible, even if they are less elegant than the phrases used in common discourse.

The word "initially" is important because it emphasizes that children are being differentiated on the basis of something measured at a single point in time rather than something fixed. The words "performing" or "achieving" emphasize that any test measures the ability to do something specific, and does not even aim to capture all mental capacities. As Feinstein (2015) and Goldstein (1979) among others have argued, it is neither necessary nor appropriate to insist that a single uni-dimensional construct be used to characterise children's development over long periods of time. Nor is it necessary, we would add, to seek to measure the abstract concept of "cognitive ability" via specific tests of vocabulary, or general intelligence or Key Stage mathematics skills. All of these measure something different and substantively important in terms of children's capacities, and can simply be discussed in terms of what they are. To say that a child has a poorly-developed vocabulary may be a narrower statement, but it embodies fewer assumptions and is more precise, than the claim that the child is "low ability" or "dim" (particularly if that child is from a foreign language background, for example).

The labelling by researchers of some children as able or not able implicitly supports the idea that this is a useful way to think about human development. One needs only to think about the debates that led to the abolition of the selective "11-plus" exam in large parts of England in the 1960s to see that such a view has long been politically contested. The idea that certain individuals inherently have more potential than others is one that plays a key role in fundamental debates over social and educational inequality (Dorling, 2015), and the incautious use of language by researchers can give the erroneous impression that the evidence they present is supportive a particular view.

This point is perhaps even more relevant in the light of the disproportionate focus on "highly able" low SES children, relative to the majority who are presumably "not highly able", in much of the comment and analysis that has followed from Feinstein's chart. The key message that provoked a sense of injustice in numerous policymakers and commentators was that, in the words of Jerrim and Vignoles "highly able children from disadvantaged homes are overtaken by their rich (but less able) peers before the age of 10 in terms of their cognitive skill" (e.g. Baumberg, 2011; Harford, 2012; HM Government, 2003, p.19; statement by the Deputy Prime Minister, HC Deb 5 April 2011). Several pieces of related research have then focused specifically on the outcomes of low SES children classed as "high-attaining" (Crawford, Macmillan & Vignoles, 2014), "talent[ed]" (Education Datalab, 2015), and "bright" (Sammons, Toth & Sylva, 2015) at some baseline assessment age. It is surprisingly difficult to find an explicit rationale for why the highly able sub-set of low SES children is singled out in this way, nor (where it occurs) for the purpose of comparing their outcomes with a low-ability high-SES group.

The idea that it is even meaningful to talk about the "overtaking" of an initially high-achieving group by a different low-achieving group is problematic, as we discuss in the following section, but even accepting that this phenomenon can be objectively identified it is somewhat mysterious as to why it matters. As stated in the introduction, the key question from a social equality perspective seems to us to be whether low SES children systematically underperform relative to higher-SES children with identical initial capacities. It is not clear why the significance of the underperformance of a particular group of high-achieving low-SES children should depend on the relative over-performance of a very

different group of low-achieving high-SES children, as seems implicit in so many interpretations of Feinstein's chart. It seems possible that the exclusion of any results for children with middling achievement levels at baseline from the chart may have unintentionally skewed reactions to Feinstein's findings. It is interesting to note that the US literature on the evolution of black-white achievement gaps, which addresses similar social and identical methodological issues, places no particular emphasis on high-achieving black children (e.g. Phillips, Crouse & Ralph, 1998; Reardon, 2008; McDonough, 2015). This suggests that a preoccupation with the initially high-achieving among the disadvantaged is not in some sense "natural". It is likely, of course, that there are many legitimate reasons for the slant taken in reactions to Feinstein, but when these are not articulated it risks giving the impression that the relative underperformance of the vast majority of low-SES children is viewed as matter of lesser social concern than the outcomes of an exceptional few. Ultimately, our main contention is that there is a continuum of both achievement and disadvantage that needs to be explored to inform policy, and our empirical example in this paper provides one example of how this might be done.

Statistical issues

The issue of potential measurement error bias in any estimation of cross-group differences in trajectories is one that, as Jerrim and Vignoles (2013) rightly emphasise, must be addressed. Their analysis traces out the consequences of the fact that there is a difference between the true score of an individual that a test attempts to measure, and the observed score that actually results from the test, which contains some component of random measurement error. The characterisation of this issue made by them, and a number of authors, as one of "regression to the mean" (RTM) has, as Feinstein (2015) notes, led to a lack of clarity in a number of areas. As Jerrim and Vignoles fully recognise, the correlation of outcomes over time that produces RTM is affected by many factors that are nothing to do with mismeasurement of the underlying constructs: factors such as transitory influences on development and shifts in the underlying skills that are relevant at different ages. It seems confusing, therefore, to refer to all these processes as RTM, and even more to describe RTM as a "spurious statistical artefact".

The real advantage of making this distinction and of framing the problem as one of measurement error is not just greater clarity, however. In this context it immediately makes transparent the links between this problem and a vast body of statistical results on how to tackle measurement error in a range of different forms (see e.g. Fuller, 2009; Carroll, Ruppert, Stefanski & Crainiceanu, 2006). To give an illustration, Jerrim and Vignoles propose the use of an auxiliary test score to categorize children as "high" or "low" ability on the first measurement occasion as a way to correct for measurement error. Although not acknowledged as such, this method can be viewed as the basis of a rudimentary instrumental variables (IV) estimator, an approach that has long been used as a standard solution in this context (e.g. Blackburn and Neumark, 1992; Ecob and Goldstein; 1983). The IV method is presented formally in the online supplementary material and illustrated in practice in our application in the next section. The advantage of drawing on the classic IV framework here is that it allows one to harness all the associated statistical results and software that have been developed over many years. As a second example of how results from the measurement error literature can inform our analysis of trajectories, we also employ an alternative method in the next section based on extensions to the textbook exposition of attenuation bias (e.g. Wooldridge, 2010). This method provides explicit expressions for the degree of bias in the estimates induced by a given amount of measurement error, and so enables systematic testing of the sensitivity of the results.

A statistical issue with Feinstein's chart that has received much less attention than measurement error is the use of cut-points to classify children into discrete high and low ability groups. On a simple level this specification is wasteful in terms of data because information on each child's individual test score at baseline is discarded. More problematically, it makes it impossible to identify the difference in later outcomes between children from different SES groups who were *identical* at baseline which, as argued, we believe is the ultimate quantity of interest. The problem is that if higher SES children have higher baseline achievement on average than lower SES children, even in the absence of measurement error the

achievement levels of those who make it into the "high achievement" group will be systematically greater than those of the low SES children classified in the same group. Comparisons of the later outcomes of the two groups of children will not be comparing like with like, and differences in subsequent rates of progress will be confounded by differences in initial conditions. These initial differences are potentially very large when cut-points such as the 75th percentile are imposed on the data. If we take the distributions of true baseline achievement in the population used in Jerrim and Vignoles' simulations as an example, the score of the average high SES child in the top quartile must exceed the score of the average low SES child in the top quartile by over a third of a standard deviation[1], and note that in this example there is no measurement error at all in the data and hence no misclassification.

The use of cut-points also has implications for the question of whether initially high-scoring low SES children are "overtaken" by initially low-scoring high SES children, which we highlighted previously. Visually this is represented in Feinstein's chart by the "crossing" of the relevant trajectories between the ages of five and ten. Quite apart from the question of why this crossing is worthy of particular attention, whether it occurs or not is highly likely to depend on arbitrary definitions of the high and low achievement categories. As we show in the empirical example in this paper, slight modifications to the presentation can generate or eliminate the appearance of overtaking in predictions from a single underlying model. Hence without greater precision in terms of definition, the statement that this phenomenon does, or does not, occur is essentially meaningless.

A framework that relates continuous measures of baseline achievement to later outcomes overcomes these problems, by allowing the baseline measures to be "matched" exactly across groups, and by forcing the analyst to be transparent about which comparisons have been selected for presentation and why. In contrast, one advantage of using raw group-specific means is that the way in which the results have been generated is immediately intuitive for a non-technical audience. It is possible there are other advantages and it would be helpful if these were articulated in applications where cut-offs are imposed.

Estimating SES differentials in trajectories in a US cohort

In this section we present some evidence on trajectories of relative reading achievement by SES that illustrates the potential of several analytical methods. The data used are taken from the Early Childhood Longitudinal Study - Kindergarten cohort (ECLS-K), a nationally representative longitudinal study of US children who entered kindergarten in 1998. An initial sample of around 19,000 children, along with their parents and teachers, were surveyed on six occasions between entry to kindergarten (average 5.7 years of age) and eighth grade (average 14.2 years of age). This analysis draws on the sample of 7,340 children with valid data in all six waves[2]. Longitudinal survey weights and design variables are provided by the ECLS-K to allow inferences about the underlying national population on the basis of this sample; these are used in all analyses. Reading outcomes are captured on each of the six occasions by the ECLS-K's reading theta score, derived from a suite of tests designed to measure achievement in six dimensions of reading skill, from basic letter recognition through understanding and inference to demonstrating a critical stance (see Tourangea, Nord, Sorongon, Najarian & Germino, 2009, for further details). Test scores at each age were adjusted for age-in-months at assessment and standardized to mean zero unit variance z-scores using the survey weights. We measure parental SES using the highest qualification of a parent resident with the child during the kindergarten year. We distinguish a high SES group corresponding to a parent with a bachelor's college degree or more (30% of the sample); a low SES group corresponding to no parental education beyond high school graduation (37%); and a residual medium SES group (33%).

A flexible model for investigating the extent to which trajectories diverge by SES is

$$A_{i2} = \alpha_0 + \alpha_L L_i + (\beta_0 + \beta_L L_i)A_{i1} + (\gamma_0 + \gamma_L L_i)A_{i1}^2 + u_{i2} \qquad (1)$$

Where A_{it} is the "true" or perfectly-measured achievement of child i on measurement occasion t; L_i is a dummy variable indicating membership of the low (relative to high) SES group, which can easily be extended to a vector distinguishing multiple SES categories; and u_{i2} is a mean zero uncorrelated residual term. The inclusion of the quadratic term, A_{i1}^2, allows the strength of the

association between initial and final outcomes to vary with the level of baseline achievement (so that, for example, low initial scores can be less predictive than high ones of future outcomes). The conditional SES gap at occasion two – that is, the difference in test scores predicted to open up between low and high SES children with a truly identical achievement level at time one – is given by

$$E(A_{i2}|L_i = 1, A_{i1}) - E(A_{i2}|L_i = 0, A_{i1}) = \alpha_L + \beta_L A_{i1} + \gamma_L A_{i1}^2 \qquad (2)$$

Tests of the parameters β_L and γ_L can be conducted to assess formally whether developing inequalities are more severe among those who were initially higher or lower achievers. For example, if $\beta_L < 0$, it tells us that (at least over some range) SES gaps open up more between high, rather than lower, scoring children at baseline; if $\beta_L > 0$ it is the weaker-performing low SES children who fall relatively further behind. The inclusion of the quadratic term again allows for greater flexibility, this time in where the largest gaps are estimated to appear. In our analysis of the reading trajectories of US children presented below, the hypothesis that β_L and γ_L were jointly zero could not be rejected in the vast majority of models. In order to simplify the presentation we proceed here with a more parsimonious model that imposes the constraint that the SES differential in progress is simply a constant, α_L (that is, it is the same regardless of initial achievement level) [3].

$$A_{i2} = \alpha_0 + \alpha_L L_i + \beta A_{i1} + \gamma A_{i1}^2 + u_{i2} \qquad (3)$$

The key problem highlighted by Jerrim and Vignoles (2013) is that A_{i1} and A_{i2} are not observed directly. Instead we observe imperfect test score measures, Y_{i1} and Y_{i2} that contain random measurement error components for each individual. A standard result from the statistical literature tells us that ordinary regression estimates of equation (3) will be biased when we substitute the observed error-prone variable, Y_{i1}, for the true baseline achievement measure A_{i1} [4]. One method for correcting for this bias is to use an auxiliary variable, known as an instrumental variable, to statistically "purge" Y_{i1} of its measurement error component. This "corrected" measure of baseline achievement is then used in place of the observed value in a procedure known as two-stage least squares (2SLS).

The first key requirement for this 2SLS procedure to yield correct estimates of the equation of interest is that the error components of the test scores be uncorrelated with the chosen auxiliary variable or instrument. This assumption is likely to be violated if the instrument is an alternative test taken on the same day and under the same conditions as the baseline assessment, because random environmental factors are likely to affect the two observed scores in the same way[5]. Even if this condition is satisfied, however, perhaps a more stringent requirement is that the instrument must be "redundant" in the equation of interest, that is, it contains no information about A_{i2} once true achievement A_{i1} (plus its square and L_i) are conditioned on. If the instrument has some predictive power for A_{i2} independently of the other factors included in the statistical model, then the 2SLS methods will not fully correct for measurement error biases, and may even make matters worse.

In our application we define the first measurement occasion as the spring of kindergarten, roughly half way through the child's first year of compulsory schooling (average age 6.2 years). This allows us to use a prior score on the same test taken just after kindergarten entry (about six months previously, at age 5.7 years on average) as the instrument. The outcome variables in four separate models are then reading achievement in the spring of first, third, fifth and eighth grade, or around ages seven, nine, 11 and 14 respectively. For the instrument to be valid, therefore, we must assume that measurement errors in each of the tests are independent of one another, and also that when realised (true) achievement towards the end of kindergarten is known, the observed reading test score from the start of that school year contains no further information about subsequent achievement from first grade onwards.

"Corrected" (2SLS) and "uncorrected" ordinary least squares (OLS) results for the model (3) are presented in table 1. The key parameters of interest are the coefficients corresponding to α_L, the top two rows that give estimates of the gaps predicted to appear between children who began with an identical level of reading achievement at age six, but who were from low- and medium-SES backgrounds respectively (relative to the reference high-SES group). Provided the instrumental variable assumptions are satisfied, the results in the top row show that a (marginally) significant gap of .07 standard deviations is predicted to open up by age

seven between the lowest and highest SES children who had identical reading achievement in their first year of formal schooling (12 months previously). That gap is predicted to widen steadily to .53 standard deviations by age 14. Comparison of the OLS and 2SLS results suggests that while measurement error in the test score at occasion one does indeed tend to bias naïve estimates of the size of the gaps upwards, it accounts only for around 20% of the estimated high-low SES gap in eighth grade.

Table 1. The relationships between SES, reading achievement at age six and later reading achievement

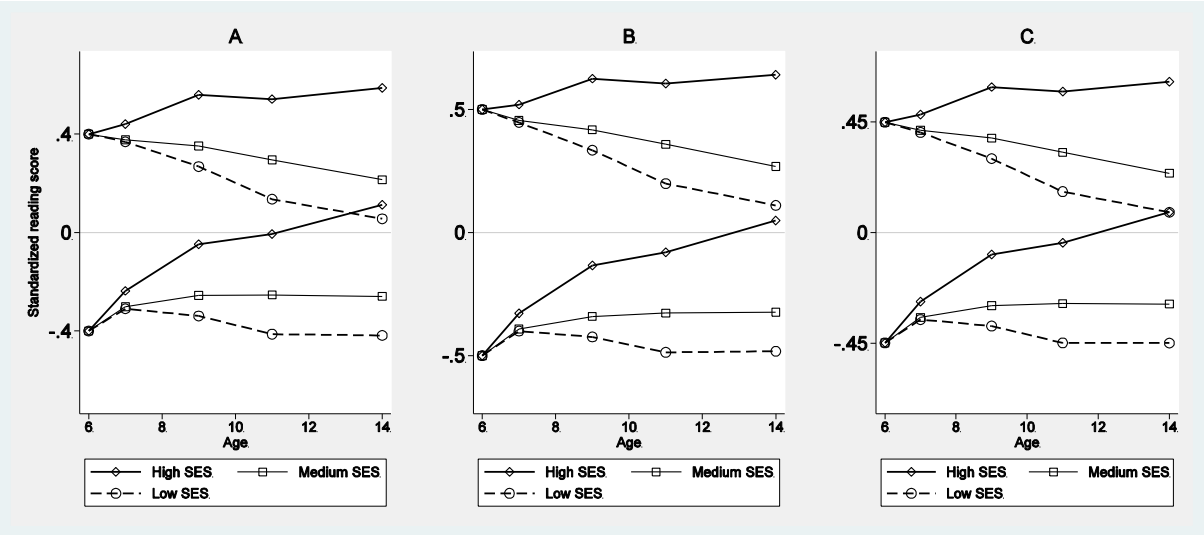
	1st grade (age 7)	3rd grade (age 9)	5th grade (age 11)	8th grade (age 14)
A. "Corrected" 2SLS estimates				
Low SES	-.073 (.037)	-.291 (.045)	-.407 (.049)	-.531 (.055)
Medium SES	-.064 (.029)	-.208 (.040)	-.247 (.041)	-.372 (.045)
Age 6 test score	.848 (.023)	.759 (.029)	.686 (.024)	.593 (.028)
Age 6 test score squared	-.067 (.014)	-.105 (.016)	-.058 (.016)	-.056 (.015)
Constant	.113 (.024)	.272 (.031)	.278 (.032)	.359 (.036)
B. "Uncorrected" OLS estimates				
Low SES	-.117 (.035)	-.443 (.044)	-.525 (.049)	-.656 (.050)
Medium SES	-.104 (.029)	-.263 (.038)	-.294 (.040)	-.417 (.044)
Age 6 test score	.735 (.015)	.595 (.021)	.557 (.019)	.458 (.020)
Age 6 test score squared	-.034 (.009)	-.043 (.011)	-.025 (.011)	-.004 (.009)
Constant	.128 (.024)	.280 (.029)	.299 (.029)	.364 (.033)

Standard errors in parentheses. High SES is the omitted reference category. N = 7340.

The results in table 1 allow a predicted outcome to be generated for any combination of SES and kindergarten test score, which can be used to illustrate the consequences of different definitions of what constitutes "high" and "low" achievement on the first measurement occasion. Panel A of figure 1 plots the trajectories predicted by the 2SLS estimates of children from different SES groups who started with achievement either .4 standard deviations above or below the mean in kindergarten. The choice of .4 as an initial benchmark generates a pattern of "crossing" trajectories: a low SES child starting from the higher achievement score is expected, on average, to have poorer outcomes by age 14 than a high SES child whose achievement was .8 standard deviations lower in the first year of schooling. Panel B plots trajectories for slightly more extreme initial achievement levels – .5 standard deviations above and below the mean respectively – and shows that with this minor difference the crossing property is eliminated. Although there is considerable convergence, a low SES child starting at the higher of the two scores is in this case predicted to continue to outperform a child who began with a one standard deviation relative disadvantage in kindergarten, even if that child is from a high SES background. To us, this illustrates the arbitrary nature of comparisons of that single out specific trajectories associated with different starting points for different socioeconomic groups. The compelling fact here is that if we take two children with identical reading achievement at age six – one from a high-SES background and one from a low-SES background – by age 14 the high-SES child is predicted to out-perform the low-SES one by over

half a standard deviation. This is equally true and, we argue, equally important, for all low-SES children, including those whose reading achievement was not exceptional when they started school.

Figure 1. Expected trajectories of reading achievement by SES, from selected starting points at age six



Note: Charts plot outcomes predicted on the basis of the 2SLS estimates in panel A of Table 1, for the selected values of the initial age 6 score marked on the y-axis.

If, however, there remains an interest in comparing initially higher-achieving low-SES children with their mirror image lower-achieving high-SES peers, then the use of continuous achievement measures, rather than arbitrary ability groupings, allows for a much more systematic analysis. For example, one could trace out the frontier of exactly how much higher the initial achievement of a low SES child than a high SES child must be in order for them to reach the same level of performance in eighth grade. This distance can be calculated for every possible outcome level in eighth grade, allowing a detailed characterisation of where "overtaking" occurs in the joint baseline achievement distribution of the two groups. Panel C of figure 1 provides just one example, showing that a high SES child scoring .45 standard deviations below the mean in kindergarten (the 33rd percentile) is predicted to attain the same reading level in eighth grade (a score of .08) as a low SES child who began .45 standard deviations above the mean (the 67th percentile).

The instrumental variables strategy outlined here is only valid if the chosen instrument satisfies the necessary statistical requirements. In our application, it also constrained our analysis of trajectories to begin with baseline achievement at age six (the second measurement occasion in the survey), because the first age five score had to serve as the auxiliary variable to correct for measurement error. In datasets where there is a long lag between the first and second measurement occasions, this approach could prevent the analysis of trajectories over important periods earlier in childhood. An alternative approach to adjusting for measurement error – the moment-based estimation method – makes a different set of assumptions, which may be more or less valid than the IV assumptions, but has the advantage that it dispenses with the need to employ an auxiliary variable at all. Instead, this approach requires that we make additional assumptions about the distributions of both unobserved true achievement and measurement error[6], and crucially that we know the degree of error in, or reliability of, the observed test score.

The reliability, denoted r, is the proportion of variance in the observed test score generated by variation in true underlying achievement. (r lies between 0 and 1, with higher values corresponding to more accurate measurement.) When r is known, correction factors can be derived and applied directly to the OLS estimates (see Goldstein, 1979

and Phillips et al., 1998 for examples of similar applications, and the supplementary material for the derivation of the correction factors applied here). Test developers often provide estimates of reliability based on the internal consistency of the individual test items, but it seems unlikely that this will capture all potential sources of noise in all contexts. In cases where the reliability is uncertain or unknown, estimates for a range of values of r can be computed to assess sensitivity. In addition, we might usefully ask how low the reliability would need to be for divergence by SES to be purely a statistical artefact.

With regard to the current application, the ECLS-K test developers provide an estimate of .95 for the internal reliability of the age six reading test score (Tourangeau et al., 2009, table 3-10). Corrected estimates using this value of the reliability are provided in table 2. Given the very high value assumed for r, it is unsurprising that these estimates are very close to the unadjusted OLS estimates shown in table 1. Alongside the coefficients on the SES indicators, table 2 also shows the minimum reliability needed to generate a negative SES gap at each age, r^*. More developed analyses could provide estimates of the reliabilities needed to generate statistically significant, rather than just non-zero, SES gaps, but it is clear that, at least from third grade onwards, observed age six test scores would have to contain an implausibly large proportion of measurement error for the finding of divergence by SES to be entirely spurious.

Table 2. Measurement-error corrected estimates of trajectories in reading achievement from age 6 onwards, *r* = .95

	1st grade (age 7)		3rd grade (age 9)		5th grade (age 11)		8th grade (age 14)	
	Coef (SE)	r^*	Coef (SE)	r^*	Coef (SE)	r^*	Coef (SE)	r^*
Low SES	-.142 (.035)	.79	-.415 (.044)	.55	-.499 (.050)	.49	-.635 (.051)	.39
Medium SES	-.088 (.029)	.75	-.249 (.038)	.50	-.282 (.040)	.46	-.407 (.044)	.32
Age 6 test score	.775 (.016)		.628 (.023)		.588 (.020)		.482 (.021)	
Age 6 test score sq	-.037 (.010)		-.048 (.012)		-.027 (.012)		-.004 (.010)	
Constant	.114 (.024)		.270 (.029)		.289 (.029)		.353 (.034)	

r^* denotes the minimum value of r consistent with a negative, non-zero estimate of the associated coefficient in a measurement-error-corrected model.

Conclusion

Feinstein's (2003) chart attracted a great deal of attention, perhaps because its strong visual image resonated with people's intuition about the way social class differences become embedded over the course of childhood. The interest generated suggests there is great value in digging deeper into precisely when and for whom trajectories diverge, and our analysis of the ECLS-K suggests a number of avenues for research that could refine our understanding in the British context.

The question of the timing of the evolution of SES gaps is important for thinking about when to target policy interventions, particularly given the recent emphasis on the preschool period as advocated by James Heckman and others. The trajectory analysis presented here suggests that educational inequality in the US strengthens considerably in the eight years after school entry, a finding that would be missed from inspection of cross-sectional achievement gaps, which change little with age[7]. The question of whether the degree to which inequalities widen varies with baseline achievement level is also one with important implications for policy. The interventions that are likely to prevent gaps emerging among initially higher-achieving children may be very different from those needed to promote equality among those struggling with basic skills. In our US cohort we find that divergence is a common problem across the full range of the initial achievement distribution, but this is unlikely to be the case in all countries and time periods. A full characterization requires that we move beyond a preoccupation with children predefined as "low" and "high"

achievers and give equal consideration to the potential of all lower SES children.

It is important that the issue of measurement error be dealt with if we are to provide convincing answers to these important questions, but we believe it is a mistake to assume any evidence of diverging trajectories must be spurious until proven otherwise. A range of methods for assessing sensitivity to and adjusting for measurement error are available. We explore only a few of these here and our results suggest that, although measurement error bias is present, it plays only a minor role in what is a much larger story about the evolution of social inequality.

Acknowledgements

This paper draws on work supported by the Russell Sage Foundation and the Australian Research Council (grant DP130103440). The authors would also like to thank Bruce Bradbury, Miles Corak, Jane Waldfogel and Anna Zhu for their help in developing the ideas in this project.

References

Baumberg, B. (2011, June 16). *The rise and fall of a killer chart* (Web log comment) Retrieved from http://inequalitiesblog.wordpress.com/2011/06/16/the-rise-and-fall-of-a-killer-chart

Blackburn, M. & Neumark. D. (1992). Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials. *The Quarterly Journal of Economics*, *107*(4), 1421-1436. http://dx.doi.org/10.2307/2118394

Bradbury, B, Corak, M, Waldfogel, J & Washbrook, E. (2015). *Too Many Children Left Behind: The US Achievement Gap in Comparative Perspective.* Russell Sage Foundation: New York (in press)

Carrol, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective, Second Edition*. London: Chapman and Hall. http://dx.doi.org/10.1201/9781420010138

Crawford, C., Macmillan, L. & Vignoles, A. (2014). *Progress made by high-attaining children from disadvantaged backgrounds.* Research report. June 2014. Social Mobility and Child Poverty Commission.

Dorling, D. (2015). *Injustice: Why social inequality still persists* (revised edition). Bristol: Policy Press

Ecob, R., & Goldstein, H. (1983). Instrumental variable methods for the estimation of test score reliability. *Journal of Educational and Behavioral Statistics*, *8*(3), 223-241. http://dx.doi.org/10.3102/10769986008003223

Education Datalab (2015). *Missing Talent*. Research brief. 5 June 2015. The Sutton Trust. http://www.suttontrust.com/wp-content/uploads/2015/06/Missing-Talent-final-june.pdf

Feinstein, L. (2003). Inequality in the early cognitive development of British children in the 1970 cohort. *Economica*, *70*, 73-97. http://dx.doi.org/10.1111/1468-0335.t01-1-00272

Feinstein (2015). Social class differences in early cognitive development and regression to the mean. *Longitudinal and Life Course Studies 6*(3)

Fuller, W. A. (2009). *Measurement error models.* John Wiley & Sons.

Goldstein, H. (1979). Some models for analysing longitudinal data on educational attainment. *Journal of the Royal Statistical Society. Series A (General)*, 407-442. http://dx.doi.org/10.2307/2982551

Harford, T. (2010, December 11). *Why education fails the poor* (Web log comment) Retrieved from http://timharford.com/2010/12/why-education-fails-the-poor/

HC Deb (2011). House of Commons Debate, 5th April 2011, c875.

HM Government (2003). *Every Child Matters*. Green Paper, September 2003, Cm 5860.

Jerrim, J. & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society series A 176*(4), 887 – 906. http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

McDonough, I. K. (2015). Dynamics of the black–white gap in academic achievement. *Economics of Education Review*, *47*, 17-33.

http://dx.doi.org/10.1016/j.econedurev.2015.03.007

Phillips, M., Crouse, J. & Ralph, J. (1998). Does the black-white test score gap widen after children enter school? Pp. 229-272. In C. Jencks, C. & Phillips, M. (Eds) *The Black-White Test Score Gap*, Washington, DC: Brookings Institution Press.

Reardon, S. F. (2008). *Thirteen ways of looking at the black-white test score gap.* Stanford Institute for Research on Education Policy & Practice, Working Paper, 8.

Sammons, P., Toth, K. & Sylva, K. (2015). *Subject to background: What promotes better achievement for bright but disadvantaged students?* Oxford University Department of Education. http://www.suttontrust.com/researcharchive/subject-to-background/

Tourangeau, K., Nord, C., Le T., Sorongon, A.G., Najarian, M. & Germino Hausken, E. (2009). *Combined User's Manual for the ECLS-K Eight-grade and K-8 Full Sample Data Files and Electronic Codebooks*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data, 2nd Edition*. MIT Press.

Endnotes

[1] In this example, baseline achievement follows a standard normal distribution and two equal-sized groups have means of -0.7 (labelled low SES) and 0.7 (labelled high SES). Average scores of children in the top quartile overall can therefore be calculated as 1.13 for the low SES group and 1.48 for the high SES group.

[2] In accordance with NCES reporting rules, all sample sizes are rounded to the nearest 10.

[3] We note, however, that the interaction terms are significant when achievement in maths, rather than reading, is the outcome of interest. See Bradbury, Corak, Waldfogel & Washbrook (2015), Ch. 6, for related analyses of these data, which include fully interacted models and estimates for both types of outcome.

[4] See supplementary material for further details.

[5] As they acknowledge, this problem is likely to affect the corrected results presented by Jerrim and Vignoles (2013).

[6] Specifically that they are normally distributed with constant variances – see the supplementary material for a formal treatment

[7] See Bradbury et al. (2015), Ch. 6 for more discussion on this point.

Commentary by

Ruth Lupton University of Manchester, UK
ruth.lupton@manchester.ac.uk

The practice of policy-related research

Introduction

There are many fascinating and important aspects of the debate over the 'Feinstein graph' which is the subject of this section. Here I address just one – what the case suggests about processes of research production, dissemination and use in public policy fields. For those reading the paper in isolation from the rest of the collection, I first offer a brief descriptive account of the case. I then explore policy implications of the findings of the Feinstein and Jerrim/Vignoles papers, both as they were put and as they were taken up. I suggest that avoidance of the wrong policy implications being drawn from this debate could be helped, among other things, by reference to research in other disciplines, which speaks to the same questions, and conclude with some questions on how such interdisciplinarity might be achieved in practice.

The debate over the Feinstein graph

In 2003, the journal *Economica* published a paper by Leon Feinstein based on analysis of the 1970 British Birth Cohort Feinstein (2003) looked at children's performance in various tests of functioning and cognitive attainment at ages 22 months, 42 months, five years and ten years. He described the average relative position in the overall rankings, at each age, of children from different social economic backgrounds (hereafter SES). This showed that at 22 months, there was already a socioeconomic gradient. This pattern persisted at 42 months, five and ten years and became more pronounced between the years of five and ten.

Feinstein then went on to group the low and high SES children according to their performance on the first test, and to look at their subsequent average rankings in later tests. On average the ranking of high attaining low SES children declined over time, while that of low attaining high SES children increased. Indeed, this latter group had overtaken the former group by age ten (i.e. during the primary school years). The visualisation of this finding (figure 2 in the original paper) is what became familiarly known as 'the Feinstein graph'. Feinstein also ran the analysis classifying groups on the basis of the second test rather than the first. The same pattern was also evident, but was less marked. The ranks of the high attaining low SES children and the low attaining high SES children were converging by age ten but the latter had not overtaken the former. A further important point, often overlooked in the debate about the converging lines, is that low SES children in the lowest quartile at 42 months made no relative gains by age ten. Their primary school experiences did not enable them to catch up with other children. Similar findings were reported in Schoon (2006) using data both from the British Cohort Study (BCS) and the 1958 National Child Development Study, and also in Goodman and Gregg (2010).

In 2011, John Jerrim and Anna Vignoles published a working paper entitled *The use (and misuse) of statistics in understanding social mobility: regression to the mean and the cognitive development of high ability children from disadvantaged homes* (Jerrim & Vignoles, 2011a) a version of which subsequently appeared in the Journal of the Royal Statistical Society with a less provocative title (Jerrim & Vignoles, 2013). This paper argued that the phenomenon of the converging trajectories of initially high and low attainers could be caused by regression to the mean. One way to correct for this is to use a different baseline measure to classify the children than the one used as the starting point of the trajectory. This is not possible in the British Cohort Study, so the authors used data from the Millennium Cohort study, with a baseline at age three. They found some convergence between groups between ages three and five (but less than found with the conventional method). Between age five and seven, the initially low attaining high SES children continued to improve their ranking, but the high attaining low SES children maintained theirs, so there was no crossover effect. For both initial high attainers and low attainers, there was some widening of the gaps between social class groups between ages three and seven.

Policy implications and policy reactions

As Feinstein recounts in his paper in this issue, the Jerrim and Vignoles working paper set off a public debate over the Conservative/Liberal Democrat Coalition Government's Social Mobility Strategy. Writing in the Guardian newspaper, Allegra Stratton suggested that Conservatives were circulating the working paper, having renamed it 'Forget Feinstein' and seeing it as a way to challenge the Social Mobility Strategy so strongly associated with the Liberal Democrat Deputy Prime Minister Nick Clegg. Quoting the paper's conclusions, she stated "in other words, the entire basis for the government's Social Mobility Strategy is wrong" (Stratton, 2011). This line was also taken by Read (2011) and by Saunders (2012). Saunders used Jerrim and Vignoles findings as one plank in an argument that Britain does not have a social mobility problem nor need a Social Mobility Strategy.

The statement that "the entire basis for the government's social mobility strategy is wrong" was a misinterpretation of Jerrim and Vignoles message, as they pointed out in a published riposte (Jerrim & Vignoles, 2011b). First of all, a careful look at the ways in which the Feinstein graph was being cited in policy documents shows that it was his evidence on the wide gaps early in life that was being used in support of policy, not the disputed point about whether and at what age initially high attaining low SES children are overtaken by initially low attaining high SES children. In the key document, the Social Mobility Strategy, the graph is reproduced in a section suggesting the need for very early intervention (Cabinet Office and Deputy Prime Minister's Office, 2011). This is also the context in which it is cited in two other influential documents: the 'Field Review', a government commissioned report on Poverty and Life Chances (Field, 2010); and the 'Marmot Review', a government commissioned report on health inequalities, where the conclusion drawn from the graph is that "even greater priority must be given to ensuring expenditure early in the developmental life cycle" (Marmot, 2010 p.22). The need for investment in the early years is supported by both papers, as Jerrim and Vignoles state explicitly in their conclusion. Their findings did not demonstrate that this element of the Social Mobility Strategy was wrong. The Coalition Government has been criticised for failing to follow through on its early years pledges, cutting funding and increasing poverty among families with the youngest children (Stewart & Obolenskaya, 2015). But that this was a result of the underlying evidence apparently being shaken is implausible, since the effective policy decisions were made prior to this debate.

So what about the issue really in dispute in the papers - whether initially high attaining children from low SES backgrounds really fall behind during the school years? Feinstein's findings on this would tend to suggest the need for policies continuing to target children from low income families during schooling and beyond. This was indeed an explicit policy of the Coalition Government, particularly through its establishment of the Pupil Premium, a new per capita funding amount for schools to spend specifically on raising the attainment of children from low-income homes. The previous Labour government had also increased the amount of funding for disadvantaged pupils and introduced a range of targeted initiatives to support their attainment (see Lupton & Obolenskaya, 2013; Lupton & Thomson, 2015 for a more extended discussion). Notably, Jerrim and Vignoles do not suggest that their work implies disinvestment in redistributive efforts after children have started school. Such policies would also be justified by the (undisputed) evidence on the persistence of wide gaps throughout childhood, unless there was evidence that there is no point (because end-of-school gaps merely reflect fixed differences evident at the start) or that there are no interventions that are effective. Feinstein points to the need for more evidence on these points in the conclusion to his paper. However this implication was taken up by Saunders (2012) who argued that if there is no evidence of wasted working class talent, there is no rationale for policies to break down barriers (for example in university admissions). Instead, Saunders argued that differences in socioeconomic outcomes are largely genetically determined and that policy makers should target the 'underclass problem', i.e. that children's lives are blighted by poor parenting in welfare-dependent households.

Again, there is no evidence that this debate actually changed policy. The government continued with its Pupil Premium and Social Mobility Strategy. However, it is clear that although not its intention, the Jerrim and Vignoles paper was used in support of arguments made in and around the Conservative Party against efforts to create more equitable

institutional processes and outcomes in the interests of social mobility, and in favour of efforts to increase the personal responsibility, resilience and skills of the poor.[i] Those arguments have arguably gained some traction. In the manifesto of the Conservative government elected in 2015 there is no mention of social mobility, nor of early intervention, early childhood disadvantage, an early years strategy, or socioeconomic inequalities in education, although the Pupil Premium policy is maintained. Those wishing to resurrect such priorities and strategies now face another round of gathering and interpreting evidence that they are needed.

How can the wrong policy implications be avoided?

Cases like this where something goes wrong in the public and political interpretation of research provide a useful if painful reminder of the difficulties academics face in working at the interface with policy. Writing from the position of someone whose findings from longitudinal research have also previously been taken up by the political right to undermine equitable policies, and as someone who works on the substantive issues discussed in these papers, I want to try to draw out some of the lessons that might be learned.

Two things appear to have gone wrong in this case. One is that public commentators failed to identify the specific challenge being made by Jerrim and Vignoles (the extent of convergence of trajectories of initially high and low attainers from different socioeconomic groups) and the policies to which it related, instead understanding that all the underlying evidence for the Social Mobility Strategy was flawed. The other is that they read Feinstein's research as the only evidence relevant to social mobility policies, thus over-stating the policy implications of a challenge to that research.

In relation to the former, the terminology of the Jerrim/Vignoles paper would appear to be implicated. Bearing in mind that the working paper and to a lesser extent the journal article are comprised principally of a dense statistical text which few people have the training to understand even if they were so minded, the choice of a controversial title to the working paper and an abstract announcing 'dramatically different results' exposing 'serious methodological problems' due to a 'spurious statistical artefact' may well have encouraged the impression that all that had gone before was now disproven. The use of terms such as 'bright' and the introduction of the notion of 'true ability', with the assumption that this is lower in working class children, may have unintentionally provided material to those who believe that intellectual capabilities are genetically fixed. Not too many years ago, such terminology used in a methodological working paper would have been inconsequential. However working papers are now published online where they are more accessible than in academic journals, and often with accompanying press releases highlighting key messages. Wide accessibility to people who need only read very few lines to see the main things that are being said substantially increases the risk that the complex analyses forming the bulk of these papers will be misinterpreted. This situation perhaps requires more than usual caution over terminology and sensitivities to the political environment in which it is taken up (not something in which academics are trained).

Both the working paper and the journal article also made a very clear link between Feinstein's work, which it challenged, and the Social Mobility Strategy. According to the article, figure 2 (the converging lines) and similar findings in later studies "has had a significant influence on both academic research and public policy in Britain" (Jerrim & Vignoles, 2013, p. 889), and it drew attention to the citations in major policy documents, and to the fact that the Deputy Prime Minister specifically cited this graph in a House of Commons Debate launching the latter strategy. Following this, the question is asked: "but is the statistical methodology lying behind this result robust?" The implication (intended or not) is that the policies were based on findings that were wrong, so it is unsurprising that this was the conclusion drawn. I make this point at length simply because this is exactly the kind of practice which is now encouraged in an environment in which researchers are measured on their ability to produce policy impact, and in which universities are keen to find newsworthy stories to promote their activities. Academic caveats, preamble and understatement are easily lost in efforts to make clear links to policy. However, it is the second of the misinterpretations that seems to me to provide the more challenging lessons. In the conclusion of their working paper, Jerrim and Vignoles sounded "a clear warning to

academics and policymakers not to place too much emphasis on one single result" (emphasis added). This was echoed by Saunders (2012, p.18), who stated that that until the publication of the Jerrim and Vignoles paper there had been "one striking and compelling piece of research" that supported politicians in their thinking that differences in social outcomes must be the result of unfair disadvantages growing up, and by David Willetts (cited in Feinstein, this issue) who suggested that policy-makers had placed too much emphasis on one result (the Feinstein graph). But how did this come to be the case, or to be seen to be the case, by a senior and well-informed government minister? In the sociology of education alone there is a vast body of work pointing to the ways in which 'environmental factors', broadly put, hold back the achievement of low SES children relative to high SES children. These include (and this is by no means a complete list), evidence on:

- differential access to high quality schools (Gewirtz, Ball, & Bowe, 1995; Reay and Lucey, 2003)
- processes of setting and streaming which serve to disproportionately allocate more disadvantaged children to lower classes (Gillborn & Youdell, 2000), and evidence that low attaining pupils are more likely to be demotivated and less likely to attain well if placed in low attaining groups (Ireson & Hallam, 2001).
- the limited pedagogies and narrowed curriculum that can arise in such classes (Thrupp, 1999; Lingard, 2007; Lupton & Hempel-Jorgensen, 2012).
- low SES children not feeling valued at school (Reay, 2006; Bright, 2011)and being demotivated by messages that they are failing.
- the effects of material poverty, housing conditions, and the social and emotional consequences of disadvantage (Ridge, 2002; Horgan, 2007).
- the effects of disadvantaged contexts on school organisation and processes, including teacher recruitment and retention (Lupton, 2006).
- social and cultural capital and the practices of middle class parents to support educational progress (Ball, 2003; Brantlinger, 2003; Lareau, 2011; Reay, Crozier, and James, 2011)

Feinstein also noted in his paper that the results would not surprise those working in the fields of developmental psychology, psychometrics and behavioural genetics. Studies in economics have also documented some of these processes and their effects, particularly school choice (e.g. Burgess, Briggs, McConnell & Slater, 2006; Allen & West, 2011). Cooper and Stewart's recent (2013) systematic review of quantitative studies finds clear evidence of the effect of material poverty on children's cognitive outcomes.

Admittedly, a more systematic review would be needed to differentiate studies that point specifically to the experiences and trajectories of initially higher and lower attaining children, and to differentiate processes in early years, primary and secondary school. Many studies are small scale. We need to know more about their generalisability. They have also been conducted in different time periods and we need to know whether things have changed over time. It may be the case that there are no other specific studies that show the exact pattern described in figure 2 of Feinstein's paper. But readers of this large body of literature would find it no surprise that low SES children who show early signs of high cognitive attainment are less able to translate that into later academic success than their higher SES peers, nor that schools fail to transform the trajectories of initially lower attaining SES pupils. Jerrim and Vignoles' results with the Millennium Cohort Study also show the accelerating performance of the higher SES initial lower attainers, compared to lower SES peers. As Francis and Mills put it (2012, p. 254) "To observe that schools reproduce social inequality is by no means novel".

Given all this, the conclusions that should be drawn are that there are substantial gaps in early attainments which are not, on average, reversed and indeed widen during the school years, and thus that the kinds of policies that need to be explored are not only early years interventions but (inter alia) the reduction of child poverty, less social stratification in access to schools, less setting and streaming, greater funding for schools serving disadvantaged areas, the development of pedagogies and curriculum which secure ongoing engagement of marginalised learners, and efforts to support families and build social and cultural capital. Jerrim and Vignoles' paper is a valuable contribution to the measurement of educational

trajectories, but it is not one that should lead to policies being overthrown or even specific policy conclusions being drawn, and they do not claim this. So how can we ensure that policy debates are not conducted at the crude level that followed the production of this paper, but in a more sophisticated way in which multiple evidence sources are utilised?

One answer would be to be clearer about the contributions of different kinds of papers and about what is expected of academics in relation to policy issues. Researchers exploring methodological issues need not necessarily be cognisant of work in other disciplines nor should they be required to reproduce all these findings in their working papers. We need disciplinary specialists and focused enquiries on specific problems. But we should not then expect or encourage disciplinary specialists to pronounce upon policy issues, which demand a wide spectrum of knowledge across disciplines. This is exactly the situation that is developing in the UK, with 'impact' playing a substantial part in the regular university research quality assessment exercises that partly determine university funding and rankings. These imperatives create a real danger that academics working within their own disciplines will be encouraged to find meaningful policy implications to add on to their scholarly articles in order to further careers and promote university reputations, although most operate without detailed knowledge of the policies upon which they are commenting or the wider evidence base that should inform these. This is a combination which will not help evidence-based policy-making.

On the other hand, if academics are expected to comment on policy issues in a particular field, for example education or housing, we might reasonably expect them to know about work in other disciplines and to be able at least to comment on the concurrences, tensions and disagreements. This has multiple implications for academic training, traditions of publication, and career advancement, all of which privilege the individual discipline, and do not incentivise this wider knowledge to develop. In particular it has implications for the ways we organise academic knowledge exchange through conferences and publication. How should an economist working with cohort study data know how to interpret her findings alongside those of findings from pyscho-social studies, organisational sociology or psychometrics? How should an ethnographically trained sociologist know what to make of longitudinal data analyses? If we expect academics to operate as policy experts, then we must find ways of creating dialogue between researchers and a shared body of knowledge, and ways to synthesise and communicate findings across disciplines.

This case also questions the degree of effectiveness of communication of research findings to the civil service. It is noticeable that the last government's Social Mobility Strategy cited not a single one of the wider evidence sources referred to in this paper – preferring to rely on a narrow range of sources (almost exclusively quantitative and mainly economic). This may reflect the disciplinary knowledge or prejudices of the people concerned, perhaps a view (mistaken in my opinion) that the only valid kind of evidence is quantitative, and with a preference for studies that establish causality (Spicker, 2011). It may also be because other disciplines are less effective in communicating to policy audiences. Working papers and policy briefings are much less common in predominantly qualitative disciplines than they are in quantitative ones, and little academic priority is given to the synthesis of existing small-scale studies, published in books and journals, so that policy-makers can see what they collectively add up to.

I do not claim to have the answers to these questions, and this short paper does not offer room to cover them, even if I did. I hope merely to have raised some of issues and highlighted the need to address them collectively. The debate over the Feinstein graph has illuminated not only the complexities of measurement of educational trajectories but also the complexities of academic practice in a changing environment.

References

Allen, R. & West, A. (2011). Why Do Faith Secondary Schools Have Advantaged Intakes? The Relative Importance of Neighbourhood Characteristics, Social Background and Religious Identification amongst Parents. *British Educational Research Journal 37* (4): 691–712. http://dx.doi.org/10.1080/01411926.2010.489145

Ball, S.J. (2003). *Class Strategies and the Education Market the Middle Classes and Social Advantage.* London: RoutledgeFalmer. http://dx.doi.org/10.4324/9780203218952

Brantlinger, E. (2003). *Dividing Classes: How the Middle Class Negotiates and Rationalizes School Advantage*. Taylor & Francis. http://dx.doi.org/10.4324/9780203465479

Bright, G. N. (2011). 'Off The Model': Resistant spaces, school disaffection and 'aspiration' in a former coal-mining community. *Children's Geographies 9* (1) http://dx.doi.org/10.1080/14733285.2011.540440

Burgess, S., Briggs, A., McConnell, B., & Slater, H. (2006). *School Choice in England: Background Facts. Working Paper 06/159*. Bristol: Centre for Market and Public Organisation.

Cabinet Office and Deputy Prime Minister's Office (2011). *Opening Doors, Breaking Barriers: A Strategy for Social Mobility*. 26. London: HM Government. https://www.gov.uk/government/publications/opening-doors-breaking-barriers-a-strategy-for-social-mobility.

Centre for Social Justice. (2007). *Breakthrough Britain: Ending the Costs of Social Breakdown*. London: Centre for Social Justice.

Cooper, K. & Stewart, K. (2013). D*oes Money Affect Children's Outcomes: A Systematic Review*. York: Joseph Rowntree Foundation.

DFE& DWP. (2012). *Measuring Child Poverty: A Consultation on Better Measures of Child Poverty*. London: Department for Education.

Feinstein, L. (2003). Inequality in the Early Cognitive Development of British Children in the 1970 Cohort. *Economica 70*, 73–97. http://dx.doi.org/10.1111/1468-0335.t01-1-00272

Field, F. (2010). *The Foundation Years: Preventing Poor Children Becoming Poor Adults: The Report of the Independent Review on Poverty and Life Chances.* London: HM Government.

Francis, B. & Mills, M. (2012). Schools as Damaging Organisations: Instigating a Dialogue Concerning Alternative Models of Schooling. *Pedagogy, Culture & Society 20* (2), 251-71. http://dx.doi.org/10.1080/14681366.2012.688765

Gewirtz, S., Ball, S.J. & Bowe, R.. (1995). *Markets, Choice and Equity in Education*. Buckingham ; Philadelphia: Open University Press.

Gillborn, D., & Youdell, D. (20000. *Rationing Education: Policy, Practice, Reform and Equity*. Buckingham England: Philadelphia: Open University Press.

Goodman, A., & Gregg, P. (2010). *Poorer Children's Educational Attainment: how Important Are Attitudes and Behaviour?*. York: Joseph Rowntree Foundation.

Horgan, G. (2007). *The Impact of Poverty on Young Children's Experience of School.* York: Joseph Rowntree Foundation.

Ireson, J., & Hallam, S. (2001). *Ability Grouping in Education*. London: Sage.

Jerrim, J, & Vignoles, A. (2011a). The Use (and Misuse) of Statistics in Understanding Social Mobility: Regression to the Mean and the Cognitive Development of High Ability Children from Disadvantaged Homes. *DoQSS Working Paper 11-01*. Department of Quantitative Social Science - UCL Institute of Education, University College London. Retrieved from https://ideas.repec.org/p/qss/dqsswp/1101.html.

Jerrim, J, & Vignoles, A. (2011b, April 28). Response: Our Early-Years Research Does Not Contradict the Government. *The Guardian*. Retrieved from http://www.theguardian.com/commentisfree/2011/apr/28/social-mobility-early-years.

Jerrim, J. & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society series A 176*(4), 887 – 906.
http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x
Lareau, A. (2011). *Unequal Childhoods*. 2nd ed. Oakland: University of California Press. http://www.ucpress.edu/book.php?isbn=9780520271425.
Lingard, B. 2007. Pedagogies of Indifference. *International Journal of Inclusive Education 11* (3), 245–66.
http://dx.doi.org/10.1080/13603110701237498
Lupton, R. (2006). Schools in Disadvantaged Areas: Low Attainment and a Contextualised Policy Response. In. Lauder, H., Dillabough, J. & Halsey, A.H. (Eds) *Education, Globalization and Social Change.* 654–72. Oxford: Oxford University Press.
Lupton, R., & Hempel-Jorgensen, A. (2012). The Importance of Teaching: Pedagogical Constraints and Possibilities in Working-Class Schools. *Journal of Education Policy 27* (5), 601–20.
http://dx.doi.org/10.1080/02680939.2012.710016
Lupton, R., & Obolenskaya, P. (2013). Labour's Record on Education: Policy, Spending and Outcomes 1997-2010. *Social Policy in a Cold Climate Working Paper WP03.* London: CASE, LSE.
Lupton, R., & Thomson, S. (2015). The Coalition's Record on Schools: Policy, Spending and Outcomes 2010-2015. S*ocial Policy in a Cold Climate Working Paper 13*. London: Centre for Analysis of Social Exclusion,.
Marmot, M. G. (2010). *Fair Society, Healthy Lives: The Marmot Review ; Strategic Review of Health Inequalities in England Post-2010*. [London].
Read, D. (2011, April 11). *Regression to the Mean: Researcher warns that Government Strategy for Social Mobility misled by a statistical trap*. Warwick University press release. Retrieved from http://www2.warwick.ac.uk/newsandevents/pressreleases/extreme_statistics/detail/.
Reay, D. (2006). The Zombie Stalking English Schools: Social Class and Educational Inequality. *British Journal of Educational Studies 54* (3), 288-307.
http://dx.doi.org/10.1111/j.1467-8527.2006.00351.x
Reay, D., Crozier, G. & James, D. (2011). *White Middle Class Identities and Urban Schooling*. Basingstoke: Palgrave Macmillan. http://www.palgraveconnect.com/pc/doifinder/10.1057/9780230302501.
http://dx.doi.org/10.1057/9780230302501
Reay, D., & Lucey, H. (2003). The Limits of "Choice": Children and Inner City Schooling. *Sociology: The Journal of the British Sociological Association 37* (1),121–42.
http://dx.doi.org/10.1177/0038038503037001389
Ridge, T. (2002). *Childhood Poverty and Social Exclusion: From a Child's Perspective*. Bristol: Policy Press. http://opus.bath.ac.uk/1294/.
Saunders, P. (2012). *Social Mobility Delusions*. London: Civitas.
Schoon, I. (2006). *Risk and Resilience: Adaptations in Changing Times*. Cambridge: Cambridge University Press.
http://dx.doi.org/10.1017/CBO9780511490132
Spicker, P. (2011). Generalisation and Phronesis: Rethinking the Methodology of Social Policy. *Journal of Social Policy 40* (01), 1-19.
http://dx.doi.org/10.1017/S0047279410000334
doi:10.1017/S0047279410000334.
Stewart, K. & Polina Obolenskaya, P. (2015). The Coalition's Record on the Under Fives: Policy, Spending and Outcomes 2010-2015. *Social Policy in a Cold Climate Working Paper WP12.* London: Centre for Analysis of Social Exclusion, LSE.
Stratton, A. (2011, April 14). David Davis Takes up Challenge to Prepare next Round of Tory Policies'. *The Guardian*. Retrieved from http://www.theguardian.com/politics/2011/apr/14/david-davis-challenge-tory-policies.
Thrupp, M. (1999). *Schools Making a Difference: Lets Be Realistic!*. First Edition. Buckingham England ; Philadelphia: Open University Press.

Endnotes

[i] These arguments are also prominent in Conservative documents on the causes of child poverty (DFE and DWP, 2012) , and in the thinking of the Conservative think-tank the Centre for Social Justice, set up by the Work and Pensions Minister Iain Duncan-Smith (Centre for Social Justice, 2007).

Referencing

The debate should be referenced as: Feinstein, L., Jerrim, J. & Vignoles, A., Goldstein, H. & French, R., Washbrook, E. & Lee, R. & Lupton, R. (2015). Social class differences in early cognitive development debate. *Longitudinal and Life Course Studies, 6,* 331-376

The individual contributions may be referenced as:
[Authors(s) names], [title of contribution], in Feinstein, L., Jerrim, J. & Vignoles, A., Goldstein, H. & French, R., Washbrook, E. & Lee, R. & Lupton, R. (2015). Social class differences in early cognitive development debate. *Longitudinal and Life Course Studies, 6, 331-376*